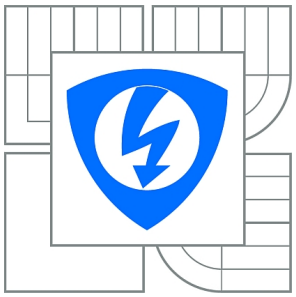
VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

BRNO UNIVERSITY OF TECHNOLOGY



FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ

ÚSTAV BIOMEDICÍNSKÉHO INŽENÝRSTVÍ

FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION

DEPARTMENT OF BIOMEDICAL ENGINEERING

ČÍSLICOVÉ ZPRACOVÁNÍ ROSTLINNÝCH GENOMŮ

DIGITAL PROCESSING OF PLANT GENOMES

BAKALÁŘSKÁ PRÁCE

BACHELOR'S THESIS

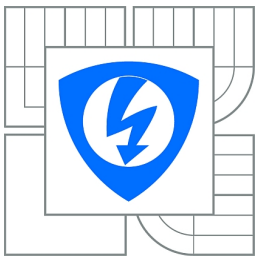
AUTOR PRÁCE — ROBIN JUGAS

AUTHOR

VEDOUCÍ PRÁCE — Ing. KAREL SEDLÁŘ

SUPERVISOR

BRNO 2014



VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

Fakulta elektrotechniky a komunikačních technologií

Ústav biomedicínského inženýrství

Bakalářská práce

bakalářský studijní obor
Biomedicínská technika a bioinformatika

Student:	Robin Jugas	***ID:***	147488
Ročník:	3	***Akademický rok:***	2013/2014

NÁZEV TÉMATU:

Číslicové zpracování rostlinných genomů

POKYNY PRO VYPRACOVÁNÍ:

1) Zpracujte literární rešerši metod konverze konvenčního znakového zápisu DNA do signálové podoby. Zaměřte se na 1D signálové reprezentace s vývojem podél genomické sekvence. 2) Popište vlastnosti a rozdíly genetického kódu jaderné a mitochondriální DNA se zaměřením na rostlinnou říši. 3) Zhodnoťte využitelnost vybraných signálů při klasifikaci organismů. Proveďte ukázku na reálných sekvencích získaných z veřejných databází. 4) V programovém prostředí MATLAB vytvořte aplikaci s grafickým rozhraním s funkcí konverze genomických sekvencí do několika různých signálových reprezentací. Aplikaci doplňte o možnost klasifikace skupiny zpracovávaných sekvencí. 5) Pomocí vytvořené aplikace porovnejte různé metody konverze z hlediska klasifikace získaných signálů a výsledky diskutujte.

DOPORUČENÁ LITERATURA:

[1] CRISTEA, P. D., Conversion of nucleotides sequences into genomic signals. Journal of Cellular and Molecular Medicine, 2002, 6(2), 279–303.
[2] ANASTASSIOU, D. Genomic signal processing. Signal Processing Magazine, IEEE, 2001.

Termín zadání:	10.2.2014	***Termín odevzdání:***	30.5.2014

Vedoucí práce: Ing. Karel Sedlář
Konzultanti bakalářské práce:

prof. Ing. Ivo Provazník, Ph.D.
Předseda oborové rady

UPOZORNĚNÍ:

ABSTRAKT

Práce navazuje na rozvoj v oblasti numerických reprezentací DNA sekvencí v uplynulých letech. Cílem bakalářské práce je zpracovat přehled numerických reprezentací DNA sekvencí a popsat vlastnosti a rozdíly genetického kódu jaderného a mitochondriálního genomu se zaměřením na rostliny. Zakončením je zhodnocení využitelnosti daných signálových reprezentací pro klasifikaci organismů. Teoretická část se zabývá popisem biologických skutečností, přehledem metod konverze DNA sekvence do signálové podoby, metodami klasifikace organismů a algoritmem DTW. Praktická část sestává z vytvoření uživatelské aplikace pro klasifikaci organismů na základě numerických reprezentací a analýza využitelnosti těchto reprezentací pro klasifikaci. Výstupy shlukové analýzy numerických sekvencí jsou srovnány s fylogenetickým stromem.

KLÍČOVÁ SLOVA

bioinformatika, mitochondrie, DNA, DNA signály, DTW

ABSTRACT

This work continues in development of DNA numerical representation's field in the recent years. The aim of this bachelor thesis is to work out an overview of numerical representations of DNA sequences and to describe the differences and properties of nuclear and mitochondrial genetic code focused on plants. Final objective is analysis of usability these signal's representations for classification of organisms. The theoretical part is focused on description of biological facts, overview of conversion methods of DNA sequences into signals, the methods of organisms classification and the DTW algorithm. The practical part contain the created GUI application for organism classification based on numerical sequences and the analysis of usability these numerical representations for classification. The outputs of cluster analysis of numerical sequences are compared with the phylogenetic tree.

KEYWORDS

bioinformatics, mitochondrion, DNA, DNA signals, DTW

JUGAS, Robin *Číslicové zpracování rostlinných genomů*: bakalářská práce. Brno: Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, Ústav biomedicínského inženýrství, 2014. 68 s. Vedoucí práce byl Ing. Karel Sedlář

PROHLÁŠENÍ

Prohlašuji, že svou bakalářskou práci na téma „Číslicové zpracování rostlinných genomů“ jsem vypracoval samostatně pod vedením vedoucího bakalářské práce a s použitím odborné literatury a dalších informačních zdrojů, které jsou všechny citovány v práci a uvedeny v seznamu literatury na konci práce.

Jako autor uvedené bakalářské práce dále prohlašuji, že v souvislosti s vytvořením této bakalářské práce jsem neporušil autorská práva třetích osob, zejména jsem nezasáhl nedovoleným způsobem do cizích autorských práv osobnostních a/nebo majetkových a jsem si plně vědom následků porušení ustanovení § 11 a následujících autorského zákona č. 121/2000 Sb., o právu autorském, o právech souvisejících s právem autorským a o změně některých zákonů (autorský zákon), ve znění pozdějších předpisů, včetně možných trestněprávních důsledků vyplývajících z ustanovení části druhé, hlavy VI. díl 4 Trestního zákoníku č. 40/2009 Sb.

Brno ....

(podpis autora)

PODĚKOVÁNÍ

Rád bych poděkoval vedoucímu bakalářské práce panu Ing. Karlu Sedlářovi za odborné vedení, konzultace, trpělivost a podnětné návrhy k práci.

Brno ....

(podpis autora)

OBSAH

SEZNAM OBRÁZKŮ

SEZNAM TABULEK

ÚVOD

V důsledku rozvoje sekvenovacích technik došlo k usnadnění a urychlení procesu sekvenace DNA živých organismů včetně lidského genomu. Databáze bioinformatických dat se rozrostly o sekvence mnohých organismů a umožnily tak rozvoj bioinformatiky jako takové. Paletu standardních metod bioinformatiky můžeme rozšířit o metody z oblasti zpracování číslicových signálů, jestliže dokážeme převést znakovou posloupnost DNA sekvence do číslicové podoby. Těchto metod určených jak k další aplikaci metod číslicového zpracování signálů, tak k přímé vizualizaci DNA sekvence značně přibylo. Liší se i svým konkétním využitím.

Na začátku práce bych chtěl popsat a nastudovat nezbytné biologické aspekty tématu. Jde o popis stavby deoxyribonukleových kyselin od chemických základů po výslednou podobu dvoušroubovice DNA. Chtěl bych popsat pojmy jako genom a genetický kód, popsat odlišnosti mezi prokaryotním a eukaryotním chromosomem. Dále chci nastudovat popis mitochondrií, obzvláště odlišnosti jejich genetického kódu a struktury DNA oproti jaderným strukturám případně i plastidovým strukturám. Za zajímavé bych považoval popsání teorie vzniku mitochondrií a příčiny jejich výsadnějšího postavení mezi dalšími buněčnými organelami. Dále bych nastínil rozdíly mezi živočišným a rostlinným mitochondriálním genomem.

Druhou část práce bych věnoval numerickým reprezentacím sekvencí DNA. Numerické reprezentace jsou metody konverze znakové posloupnosti sekvence DNA do podoby numerické posloupnosti, kterou můžeme dále zpracovat (například technikami číslicového zpracování signálů) či přímo vizualizovat pro zvýraznění některých rysů genetického kódu. Technik je velké množství, jsou zaměřené na různé konkrétní účely, mají své výhody a nevýhody a liší se i jejich ztrátovost a dimenzionalita. Chtěl bych je popsat a pokusit se setřídit a provést reálnou ukázku na vybraných sekvencích s použitím vlastního programového kódu.

Třetí částí práce bude teoretické uvedení do problematiky použité v analýze, která má nastínit rozdíly mezi jednotlivými numerickými reprezentacemi a klasifikací dendrogramy z nich sestrojených a ukázat na vhodnost či nevhodnost numerických sekvencí DNA pro klasifikaci. Jde o metody zarovnání znakových sekvencí, shluková analýza a metoda dynamického borcení času.

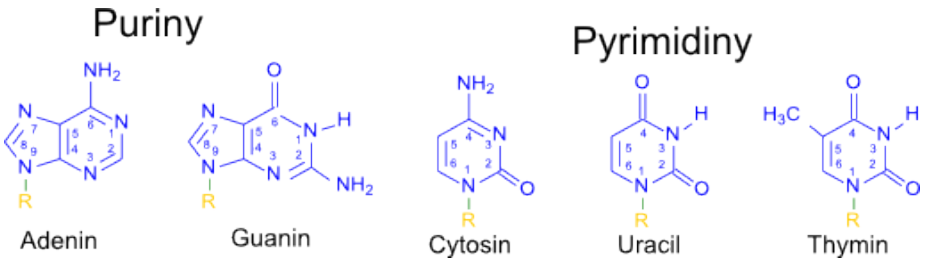
Finální částí bude praktická analýza využitelnosti těchto numerických reprezentací pro další práci – klasifikaci organismů. Získané numerické sekvence DNA bych chtěl pomocí shlukové analýzy klasifikovat a srovnat se skutečným zařazením organismů. Získal bych tak informace o tom, které reprezentace jsou vhodné pro klasifikaci v další části práce a jsou dostatečně konkrétní pro danou sekvenci. Cílem je též porovnat dendrogramy vytvořené ze znakové sekvence s dendrogramy vzniklými na základě numerické sekvence DNA.

1 NUKLEOVÉ KYSELINY

1.1 Nukleové kyseliny

Nukleové kyseliny, spolu se sacharidy a bílkovinami, patří mezi chemickou skupinu biopolymerů. Tyto látky využívají tzv. stavebnicovému principu, tedy složitější struktury (polymery) jsou tvořené z menších jednodušších monomerů. Počet těchto stavebních monomerů je vzhledem k počtu potenciálně utvořených polymerů nízký, u nukleových kyselin existuje pět základních monomerů (adenin, tymin, uracil, cytozin, guanin) a nazývají se nukleotidy. [30]

Nukleotidy jsou chemicky tvořeny ze sacharidu pentózy, na nějž je navázána dusíkatá organická zásada - baze. Na třetí atom pentózy se váže kyselina fosforečná. Sacharid pentóza může být dvojího druhu, ribóza u RNA nebo deoxyribóza u DNA. Dále známe pět dusíkatých bazí, jež dělíme na puriny (adenin A, guanin G) a pyrimidiny (tymin T, cytozin C, uracil U), jejichž chemické strukturní vzorce jsou na obrázku 1.1. RNA je kódována adeninem, guaninem, cytozinem a uracilem, DNA pak je místo uracilu kódována tyminem. [30]



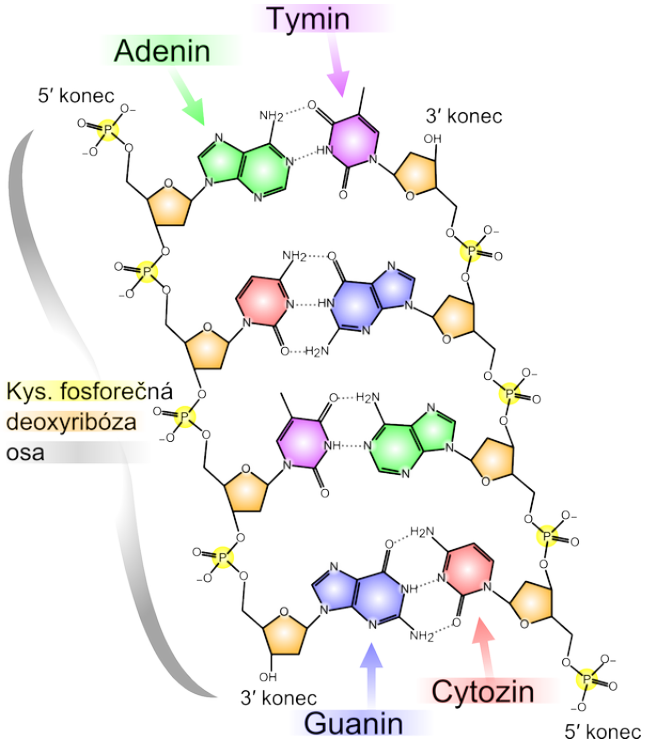
Obr. 1.1: Přehled nukleotidů (Zdroj: Wikipedia)

Nukleotidy jsou navzájem spojeny do lineárního řetězce vazbou mezi kyselinou fosforečnou na jednom nukleotidu a pátým atomem pentózy druhého nukleotidu, osu polynukleotidového řetězce tak tvoří střídající se kyselina fosforečná a monosacharid pentóza. Řetězec má tedy dva různé konce, jeden označený jako 5' konec končící fosfátem a druhý 3' konec končící pentózou. [30]

1.2 DNA

Primární struktura molekuly DNA je tvořena dvěma polynukleotidovými řetězci, jež probíhají antiparalelně, na obou koncích vlákna DNA jsou vždy 3' a 5' konce polynukleotidového řetězce. Oba polypetidové řetězce jsou spojeny vodíkovými můstky mezi bazemi, uspořádanými dle komplementarity bazí. Adenin se váže na tymin,

guanin se váže na cytozin a naopak. Z tohoto vyplývá, že počty výskytu komplementárních bazí jsou rovny. Pro bioinformatiku je stěžejní sekvence bazí (A, C, G, T) v molekule DNA, tzv. univerzální genetický kód. Molekula DNA je znázorněna na obrázku 1.2.



Obr. 1.2: Molekula DNA (Zdroj: Wikipedia)

Sekundární struktura byla popsána J. D. Watsonem a F. H. C. Crickem v roce 1953 [42]. DNA popsali jako stočenou dvoušroubovici. Nejčastější je pravotočivá dvoušroubovice s dvěma formami: B-formou a A-formou. Tato konformace je stabilizována vodíkovými vazbami mezi bazemi a vzniká samostatně, jako stav s nejmenší volnou energií. [30]

1.3 Genetický kód

Genetický kód je podkladem vyjádření genotypu do fenotypu, tedy souhrnu vnějších znaků individua. Proces se děje postupně dle centrálního dogmatu molekulární biologie [9]. Část sekvence DNA je přepsána transkripcí do mRNA, jež dále směřuje do organely ribosomu, kde se procesem translace syntetizuje z informace v mRNA do sledu aminokyselin – primární struktury výsledného proteinu. Díky znalosti 20 aminokyselin se zprvu matematicky odhadlo, že kódovat jednu aminokyselinu může nejméně sekvence tří nukleotidů. Trojice nukleotidů (tzv. triplet) může

kódovat $4^3 = 64$ různých proteinů. Toto se později ověřilo experimentálně pokusy F.Cricka. [30]

Pořadí jednotlivých aminokyselin v proteinu je určeno triplety v molekule mRNA, proto tabulka kódování aminokyelin IUPAC obsahuje místo tyminu uracil. Kódování je tzv. degenerativní, neumožňuje zpětné rozluštění, neboť některé aminokyseliny jsou kódovány vícero kodony. Molekula mRNA skutečně využívá všech 64 kodonů, 61 z nich kóduje aminokyseliny, 3 triplety slouží k terminaci translace (terminační kodony). Současně tvoří hranice mezi jednotlivými geny v RNA. Další vědeckou prací bylo potvrzeno, že až na sporadické výjimky (např. mitochondriální DNA) je genetický kód univerzální. [30]

Proteosyntetický aparát, jež tvoří ze sledu tripletů polypeptidový řetězec je tvořen mRNA, jež nese informace o primární struktuře polypeptidu, jak byla zapsána v DNA, a tRNA, jež provádí syntézu proteinu. tRNA je složena z dané aminokyseliny a antikodonu, který je komplementární ke kodonu v mRNA. Buňka tvoří 61 různých tRNA molekul k odpovídajícím kodonům.

Obr. 1.3: Centrální dogma molekulární biologie

1.4 Genom

Deoxyribonukleová kyselina je chemickou podstatou buněčné paměti a dědičnosti. Nukleové kyseliny nesou nezbytné informace o struktuře, funkci a reprodukci buněk. Gen je dle definice úsek polynukleotidového řetězce, nesoucí informaci pro strukturu translačního produktu (strukturní gen) nebo informaci pro strukturu RNA, jež dále nepodléhá translaci. Genom je dále kompletní genetická informace nebo celá sekvence DNA konkrétního organismu.

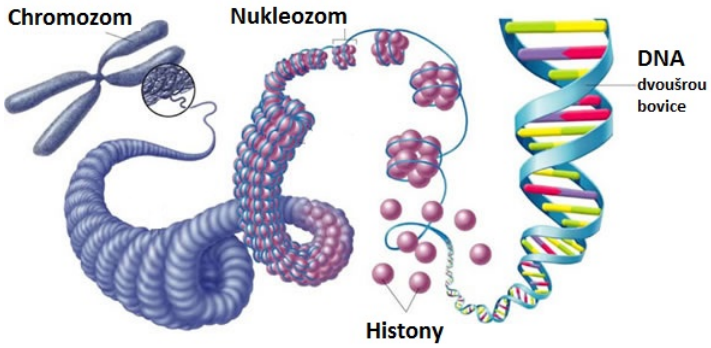
U prokaryot slouží většina DNA k genové expresi, u eukaryont má naopak většina DNA negenovou funkci. U savců jen 7 – 10 % DNA slouží k expresi do polypetidu. Tzv. negenová DNA se podílí na regulaci, ale její význam není zcela popsaný. Obsahuje však náhodné opakující se sekvence, jež mohou posloužit k identifikaci nositele.[30]

Ani genová část DNA není zcela spojitě kódující, obsahuje tzv. introny, které se nepodílí na genové funkci. Kódující sekvence se označují jako exony. [30]

Struktury, jež v buňce nesou genetickou informaci se označují jako genofory. U prokayryont jde o jeden chromosom a plazmidy, u eukaryont jsou to jaderné chro-

mosomy, mimojaderné chromosomy (mitochondrie, chloroplasty) a zřídka i plazmidy. Jaderné chromosomy označujeme jako jaderný genom, mimojaderné genofory jako plazmon či mimojaderný (cytoplazmatický) genom.[30]

Eukaryontní chromosom je tvořen DNA navázanou na bílkoviny histony. Chromosom má složitou několikaúrovňovou strukturu, začínající DNA omotanou okolo diskovitých nukleosomů, jejichž rětízky tvoří chromatinová vlákna. Ty tvoří ve vyšší struktuře smyčky, ze kterých je utvořen chromosom. Lidský chromosom má okolo 2600 smyček. Gen, kódující daný protein či RNA je umístěn na určitém chromosomu a na svém stálém místě – lokusu. Zjištěním lokusů vznikne chromosová mapa, u eukaryot lineární, u prokaryot pak kruhová. Chromosom prokaryot je totiž tvořen kružnicovou molekulou DNA. [30]



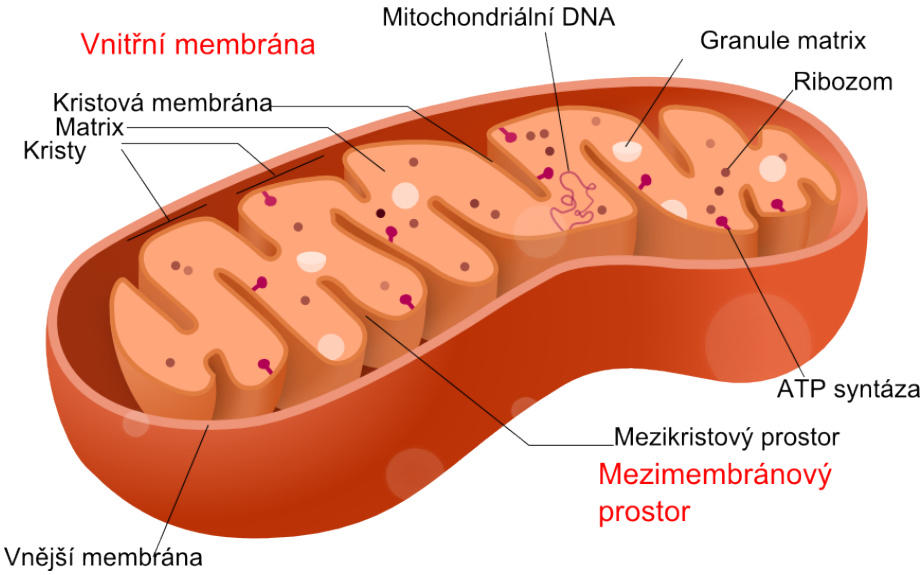
Obr. 1.4: Eukaryotní chromosom - uspořádání (Zdroj: < <www.goldiesroom.org/Multimedia/Bio_Images>>)

2 MITOCHONDRIE

2.1 Organela mitochondrie

Mitochondrie spolu s plastidy patří mezi eukaryotické organely se zvláštním postavením (tzv. semiautonomní organely), neboť obsahují vlastní DNA, proteosyntetický aparát prokaryotického typu a jsou odděleny cytoplazmatickou membránou od okolní cytoplazmy. Hlavní funkcí mitochondrií je přeměna energie z uložených zdrojů ve formě tuků, sacharidů a proteinů na energii ATP. Probíhá zde řada chemických procesů buněčného dýchání jako Krebsův cyklus, dýchací řetězec a beta-oxidace mastných kyselin.

Mitochondrie je složena ze dvou membrán, vnější jež odděluje organelu od okolní cytoplazmy a vnitřní, jež tvoří uvnitř organely záhyby – kristy. Na nich se nachází enzymy. Mezi oběma membránami je intermembránový prostor, vnitřní prostor mitochondrie je nazýván matrix. V buňce jsou často největšími organelami. [30]



Obr. 2.1: Průřez mitochondrií (Zdroj: Wikiskripta)

2.2 Endosymbiotická teorie

Endosymbiotická teorie popisuje vývoj eukaryotických buněk a odpovídá na otázku, proč eukaryotické buňky obsahují mitochondrie a plastidy s prokaryontní DNA a proteosyntetickým aparátem prokaryontního typu. [30] Vznik mitochondrií je odhadován na dobu před 1,45 mld. let při vzniku kyslíkové atmosféry. Existují aktuálně dva rozdílné scénáře – „archezoan scenario“ a „symbiogenesis scenario“ [15]. První

(archezoan) scénář, vycházející z klasické endosymbiotické teorie, předpokládá již existující primitivní eukaryotní buňku (z říše Archezoa) s anaerobním metabolismem, jež fagocytovala aerobního prokaryonta. Eukaryotické buňky, které pohltily aerobní bakterie, získaly evoluční výhodu v nově vzniklé kyslíkové atmosféře. [28] Tento pohled má své mezery, neboť známe i anaerobní typy mitochondrií a dále nebyl žádný takový předek zjištěn. [15, 28]

Druhý scénář (symbiogenesis), vycházející z vodíkové hypotézy, předpokládá naopak prokaryotní buňku z domény Archea jako hostitele pro fakultativně anaerobního prokaryonta (schopný pracovat s nebo bez kyslíku) [28]. Tato teorie současně objasňuje vznik aerobního i anaerobního metabolismu eukaryot, ale má i nedostatky nad rámec této práce.

Didaktickým modelem endosymbiotické teorie je prvok *Paulinella chromatophora*. Ten byl objeven roku 1894 německým biologem Robertem Lauterbornem. Jde o unikát, neboť endosymbióza sinice zde proběhla evolučně později a z jiných biologických rodů sinic [18]. Z biologického pohledu tak eukaryotní prvok pohltil grammnegativní bakterii schopnou fotosyntézi. Endosymbiont, tedy chloroplastový plastid, si však ponechal některé klíčové vlastnosti sinice. Genom též vykazuje znaky sinice a nikoli plastidu [44]. Část genomu předal endosymbiont jádru, ale ponechal si například typicky bakteriální gen, podílející se na fixaci vzdušného kyslíku. Na druhou stranu o dělení organely se stará gen uložený v jádře, což může nastínit rozdíl mezi organelou a endosymbiontem, kdy o dělení organely se stará hostující organismus. [18]

2.3 Mitochondriální genom živočichů

Mitochondrie mají vlastní cirkulární molekuly DNA – tzv. mtDNA, jež představují mitochondriální chromosom prokaryotického typu. Chromosomy mitochondriální DNA jsou napojeny na membránu mitochondrie podobně jako u prokaryot. Jde o cirkulární dvoušroubovici DNA s kovalentně spojenými konci. Podle obsahu mtDNA rozlišujeme tzv. těžké vlákno (též H vlákno) s převahou guaninových bazí, od lehkého (L) vlákna s převahou cytosinu. Těžké vlákno obsahuje větší množství genů. Mitochondriální DNA se vyskytuje v mitochondrii ve více,obvykle 2–10, kopiích.[30]

Genetický kód mitochondrií je odlišný od univerzálního jaderného. Obvykle kodon UGA není stopkodonem, ale kóduje tryptofan, kodon AUA kóduje místo izoleucinu methionin. Kódování mtDNA se druhově mírně liší. Na rozdíl od živočišné, rostlinná mtDNA používá univerzální kód. [30]

Mitochondriální DNA živočichů neobsahuje introny a většina (u živočichů 92%) genomu má genovou funkci [24]. Nejčastější délka sekvence mtDNA je menší než 20

tisíc bazí a kóduje u živočichů přibližne 37 genů, jež všechny souvisí s mitochondriálním dýchacím řetězcem. Z 37 obvyklých genů 13 genů kóduje polypetidy, 22 genů kóduje transferovou RNA a 2 geny kódují ribosomové podjednotky [6].

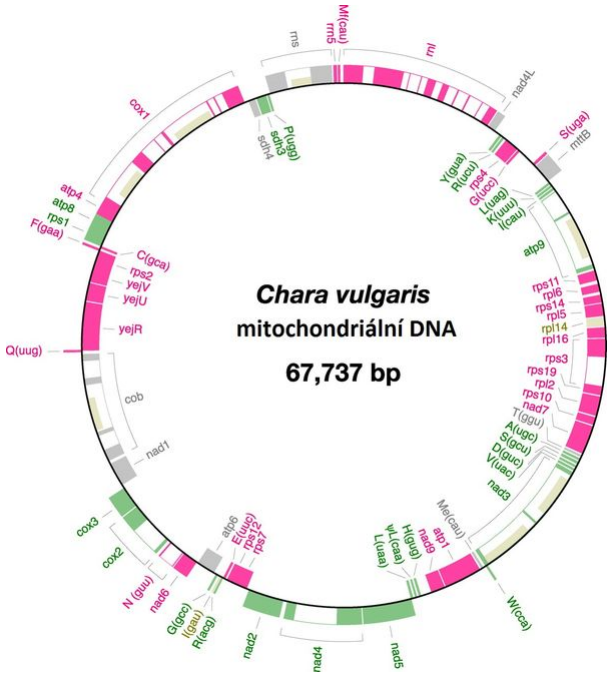
Genetika mitochondrií se od jaderné liší. Při mitóze se mitochondrie rovnoměrně rozdělí do dceřiných buněk, avšak při pohlavním rozmnožování nastává rozdíl v tom, že mateřská vajíčka obsahují velké množství cytoplazmy narozdíl od otcovských spermií, jež mají cytoplazmy, ve které se mitochondrie vyskytují, málo nebo vůbec. V důsledku se na mitochondriální dědičnosti se podílí jen matka, jde o matroklinní dědičnost. Takovou genetiku nazýváme jako nemendelovskou či cytoplazmatickou, nedochází při ní k rekombinaci. Mitochondriální genom je též polyploidní, tedy chromosomy se vyskytují ne ve dvou párech (diploide) zděděných od rodičů, ale ve vícero různých párech. [24]

2.4 Mitochondriální genom rostlin

Vyšší rostliny jsou nositely hned tří různých genetických výbav – jaderné, mitochondriální a chloroplastové. Mitochondriální genom rostlin je větší a komplexnější než u živočichů. Velikost bývá mezi 200 kbp až 2500 kbp ve srovnání se savčí mtDNA o velikostech 15 - 19 kbp [29]. Navzdory velikosti kóduje rostlinná mtDNA zhruba stejné množství genů. Narozdíl od živočišné mtDNA je kódována univerzálním genetickým kódem. Dalším rozdílem je, že většina sekvence mtDNA rostlin je považována za nekódující a také, že rostlinné genové strukturní sekvence obsahují introny. [29] Mitochondriální genom druhu *Chara vulgaris* s kódujícími geny je na obrázku 2.2.

Rychlost sekvenční a strukturální evoluce je variabilní napříč typy genomu. Míra bodových mutací mtDNA u rostlin je až 100 krát nižší než u mtDNA živočichů a 4 krát nižší než u plastidové DNA (cpDNA). Například mtDNA živočichů má neměnné uspořádání genomu, všichni obratlovci mají stejné pořadí genů. Výskyty genových přeskupení u rostlinné mtDNA jsou naopak značně vyšší oproti předchozím typům [32]. Rostlinná mtDNA cévnatých rostlin často obsahuje repetitivní sekvence, jež jsou rekombinantně aktivní a jsou zdrojem častých přeskupení. Naproti tomu plastidová cpDNA nepodléhá sekvenčním změnám ani přeskupování genových sekvencí. [39]

Díky těmto přeskupením se mitochondriální genom rostlin vyskytuje ve vícero různých strukturách, jde o tzv. „multipartite structure“. Nejčastěji se popisuje jako „tripartite“, trojdílná struktura. Metodou restrikční analýzy se dá mitochondriální chromosom namapovat do kružnic, do tzv. „master circle“, jež obsahuje celou sekvenci genomu, a do dvou menších kružnic, tzv. „subgenomic circles“, které jsou tvořeny rekombinací přes pár přímých repetitivních sekvencí [29, 39, 32]. Pokud



Obr. 2.2: Mitochondriální DNA *Chara vulgaris* (Zdroj: www.plantcell.org)

mtDNA neobsahuje žádné rekombinantní repetice vyskytuje tak se pouze jako jediný „master circle“ chromosom bez dalších „subgenomic circles“.

Kružnicové uspořádání však nemusí být nutně jediným možným a výchozím uspořádáním rostlinného mitochondriálního genomu. Studie s elektronovým mikroskopem ukazují větší výskyt lineárních molekul DNA oproti kruhovým molekulám [29]. Vyplývají z jiného pohledu na mitochondriální genom, který je spíše shlukem sekvencí s různými spojeními, než jednotným celistvým vláknem. Díky rozvoji sekvenčních technik a zaměření na mtDNA rostlin v posledních letech se tak nalézají stále častější výjimky z konvenčního uspořádání, jež nemohou být namapovány na kružnici. [39]

Například projekt sekvenování genomu *Mimulus guttatus*, česky *Kejklířka skvrnitá*. Autoři projektu uvažují spíše na genomem složeným z překrývajících se lineárních molekul DNA v nerovnoměrných výskytech[39]. „Master circle“ mtDNA též nemusí být jediným typem genomu, který by se dal namapovat do kruhu. Do kruhu se dá namapovat též kružnicový chromosom, „concatemer“ (dlouhá molekula DNA obsahující kopie stejné DNA spojené do série) či kruhově permutované lineární molekuly DNA. [39]

Díky moderním sekvenačním metodám bylo potvrzena existence multichromosomových genomů, jež se pravděpodobně vyvinuly do dvou nezávislých skupin krytosemenných rostlin. U rostliny *Cucumis sativus*, česky *Okurka setá*, byly objeveny mimo „master circle“ dva další malé chromosomy, které sdílely mezi sebou repeti-

tivní sekvence, ale s hlavním kruhem neměli společné žádné repetice [39]. Tyto dva chromosomy jsou autonomní a tvoří i organizační strukturu odlišnou od uspořádání s hlavním kruhem a podgenomickými podjednotkami nebo dodatečnými plasmidy, jež bývají též přítomny v rostlinných mitochondriích.

V genomu mtDNA se nevyskytují pouze čistě mitochondriální geny, u řady rostlin byly nalezeny v genomu sekvence homologní k plastidové (chloroplastové) cpDNA [29]. Výskyt těchto sekvencí je pravděpodobně náhodný a je výsledkem procesu, který proběhl v evolučné nedávné době [40]. Homologní sekvence byly zjištěny i mezi jadernou DNA a mitochondriální DNA, několik takových sekvencí se našlo i mezi všemi třemi typy genomů [29].

Tab. 2.1: Srovnání mtDNA živočichů a rostlin

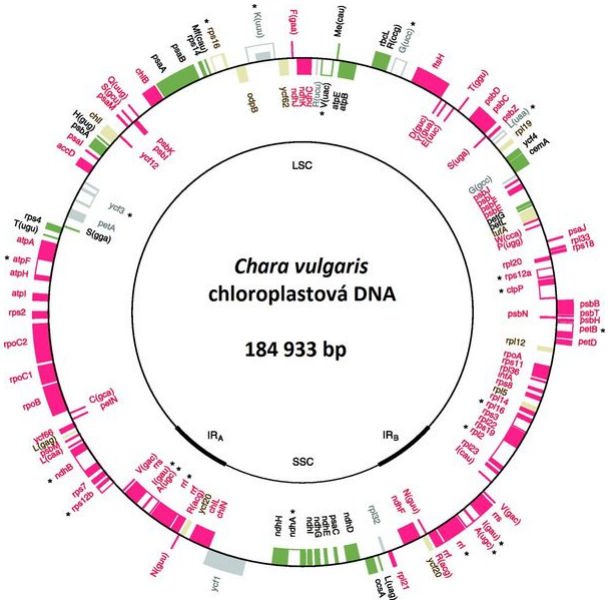
Parametr	**Živočichové**	**Rostliny**
velikost	14-42 kb	200-2400 kb
kódující proteiny	cca 13	cca 57
genet. kód	modifikovaný	univerzální
introny	ne	ano
uspořádání	cirkulární	master circle, subgenomic circles
kódující část	90 %	<10%

2.5 Chloroplastová DNA

Existence chloroplastové DNA byla poprvé prokázána roku 1962 a poprvé sekvenována roku 1986. Spolu s mitochondriální DNA patří chloroplastová DNA (cpDNA) k mimojadernému genomu. Chloroplasty existují pouze v rostlinné říši a některých bakteriích. Jejich struktura je podobná mitochondriím, mají dvě membrány, vnější a vnitřní ohraničující mezimembránový prostor a stroma. Ve stromatu jsou přítomny membránové struktury tylakoidy. Membrána tylakoidů obsahuje molekuly absorbující světlo, enzymy transportující elektrony a ATP syntézu [30]. V chloroplastech probíhají dvě hlavní reakce - fotosyntettická fosforylace (probíhající za světla) a fixace CO_2 do uhlíkatého řetezce cukrů (probíhající za tmy). [30]

Chloroplastový chromosom je prokarytoního typu, velikost dosahuje až 180 kbp. Chromosom je uložen ve stromatu ve více kopiích. Geny chloroplastového genomu kódují nukleové kyseliny proteosyntetického aparátu (rRNA, tRNA) a asi přibližně 100 proteinů spjatých převážně s fotosyntetickými pochody. U většiny rostlin nepřechází chloroplasty pylového zrna do zygoty, jde tedy o maternální nemendelovskou

dědičnost. Podobně jako mitochondrie patří chloroplasty k autoreproduktivním organelám, vznikají dělením během celého buněčného cyklu. [30] Chloroplastový genom druhu *Chara vulgaris* na obrázku **??**.



Obr. 2.3: Chloroplastová DNA *Chara vulgaris* (Zdroj: mbe.oxfordjournals.org/)

3 ČÍSLICOVÉ REPREZENTACE GENOMICKÝCH SIGNÁLŮ

3.1 Konverze nukleotidové sekvence na signál

Paleta standardních bioinformatických metod ke studiu DNA může být ještě rozšířena o metody digitálního zpracování signálů. Ty slouží k odhalení struktury genomu, nalezení periodických struktur a jiných charakteristik, které běžné bioinformatické techniky neumožní. K tomu, abychom mohli použít nástroje číslicového zpracování signálů, je třeba převést nukleotidovou sekvenci na podobu diskrétního číslicového signálu.

Konverze na číslicovou podobu DNA sekvence může být pomyslně rozdělena na numerické a grafické metody reprezentace. Grafické metody jsou vhodné k přímé vizualizaci DNA sekvence. Numerické reprezentace slouží jako převod na číslicovou podobu pro další zpracování DSP metodami, jak je znázorněno v schématu 3.1. V další kapitole jsou popsány některé numerické a dále grafické reprezentace. V různých materiálech však tyto metody naleznete pod různými názvy a rozdíly mezi nimi se stírají. [3]

Obr. 3.1: Schéma převodu na numerickou DNA sekvenci

3.2 Numerické mapování DNA sekvencí

Jak již bylo zmíněno v části o DNA, DNA je složena ze sledu nukleotidů, jimž jsou přiřazeny velká písmena A (*Adenin*), G (*Guanin*), C (*Cytosin*) a T (*Thymin*). V rámci dvojvlákna DNA se vyskytují v komplementárních dvojicích Adenin–Thymin a Guanin–Cytosin. Jednotlivým nukleotidům přiřadíme numerické vyjádření dle zvoleného typu numerické reprezentace. Typy numerické reprezentace genomických dat lze rozdělit na dvě hlavní skupiny : fixní mapování a a fyzikálně-chemicky podmíněné mapování. [23, 1]

3.3 Fixní mapování

V případě fixního mapování je sled nukleotidů převeden do podoby číselných sekvencí dle libovolných pravidel nebo matematického odvození. Mezi fixní mapování patří

Vossova reprezentace, tetrahedronová reprezentace, celočíselná, reálná a komplexní reprezentace a nakonec reprezentace krychlí.

3.3.1 Vossova reprezentace

Vossova reprezentace, jak je nazvána v některých článcích [23, 1], nebo jinak také známá jako reprezentace binárními vektory četností.[2, 11] Numerické mapování spočívá ve vytvoření čtyř vektorů pro každý nukleotid $u_A[n], u_T[n], u_G[n], u_C[n]$ dle vzorce 3.1, do kterých mapujeme sled nukleotidů logickou hodnou 1 nebo 0 podle toho, zda se odpovídající nukleotid nachází na pozici n v sekvenci DNA. [41, 2, 11]

$$u_x(n) = \begin{cases} 1, & s_n = X \\ 0, & s_n \neq X \end{cases}, X \in \{A, C, G, T\} \tag{3.1}$$

Jedná se o 4D reprezentaci, neboť každá báze je reprezentována čtyřrozměrným vektorem. Tento typ převodu na signál je velice výhodný pro spektrální analýzu DNA sekvencí a identifikace kódujících oblastí. Výkonné spektrum binárních vektorů četnosti vykazuje "peak"na frekvenci 1/3 pro kódující sekvence a pro nekódující naopak žádný výrazný "peak". [1] Ukázka takového kódování pro sekvenci ACGTA vypadá:

$$\begin{aligned} u_A(n) &= [1, 0, 0, 0, 1] \\ u_C(n) &= [0, 1, 0, 0, 0] \\ u_G(n) &= [0, 0, 1, 0, 0] \\ u_T(n) &= [0, 0, 0, 1, 0] \end{aligned} \tag{3.2}$$

3.3.2 Tetrahedronová reprezentace

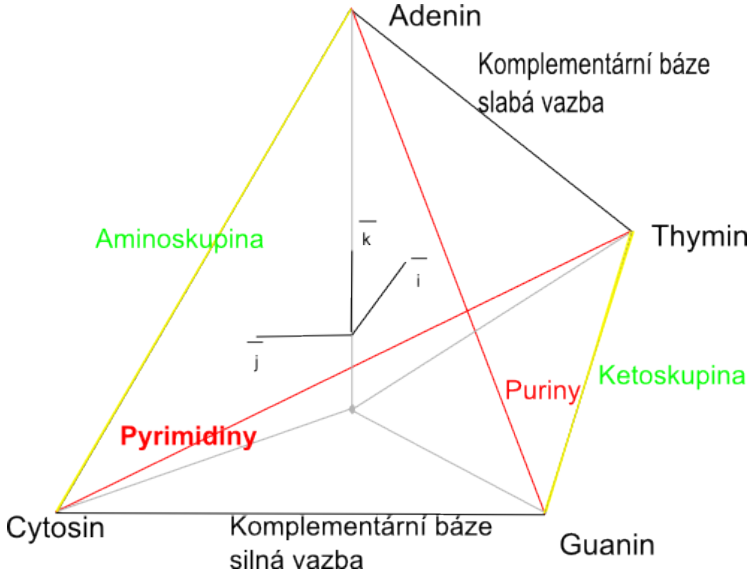
Reprezentace tetrahedronem, pravidelným čtyřstěnem, redukuje čtyřrozměrné vektory četností na trojrozměrný vektor. Každý nukleotid je přiřazen vrcholu čtyřstěnu. Čtyři binární sekvence $u_A[n], u_T[n], u_G[n], u_C[n]$ jsou převedeny ke čtyřem trojrozměrným vektorům směřujícím ze středu do vrcholů čtyřstěnu. Tyto čtyři vektory jsou zadány takto:

$$\begin{aligned} (a_r, a_g, a_b) &= (0, 0, 1) \\ (t_r, t_g, t_b) &= (\frac{2\sqrt{2}}{3}, 0, -\frac{1}{3}) \\ (g_r, g_g, g_b) &= (-\frac{\sqrt{2}}{3}, -\frac{\sqrt{6}}{3}, -\frac{1}{3}) \\ (c_r, c_g, c_b) &= (-\frac{\sqrt{2}}{3}, \frac{\sqrt{6}}{3}, -\frac{1}{3}) \end{aligned} \tag{3.3}$$

Z nich pak odvodíme samostatné vektory os:

$$\begin{aligned} x_i[n] &= \frac{\sqrt{2}}{3}(2u_T[n] - u_C[n] - u_G[n]) \\ x_j[n] &= \frac{\sqrt{6}}{3}(u_C[n] - u_G[n]) \\ x_k[n] &= \frac{1}{3}(3u_A[n] - u_T[n] - u_C[n] - u_G[n]) \end{aligned} \tag{3.4}$$

Na obrázku 3.2 je znázorněno přiřazená vektorů jednotlivým nukleotidům a souvislosti mezi jejich chemickými vlastnostmi.

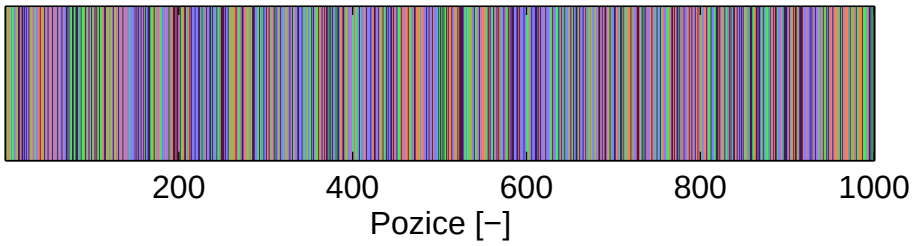


Obr. 3.2: Reprezentace tetrahedronem

Jak reprezentace binárními vektory četností, tak reprezentace čtyřstěnem jsou si ekvivalentní, což se projevuje stejnými výsledky při analýze výkonnového spektra. Čitelnost numerické sekvence při vykreslení můžeme zlepšit přiřazením základních barev systému RGB (červená, zelená, modrá) jednotlivým osám souřadnicového systému, ukázka je na obrázku 3.3. [2, 1]

3.3.3 Reprezentace krychlí

Pokud zrotujeme pravidelný čtyřstěn z předchozího případu, můžeme jej vepsat do krychle, takže vrcholy čtyřstěnu odpovídají některým vrcholům krychle. Současně můžeme zjednodušit systém vektorů přiřazených nukleotidům tak, že hraně krychle přiřadíme délku rovnou 1 a výsledné vektory přiřazené nukleotidům směřující ze

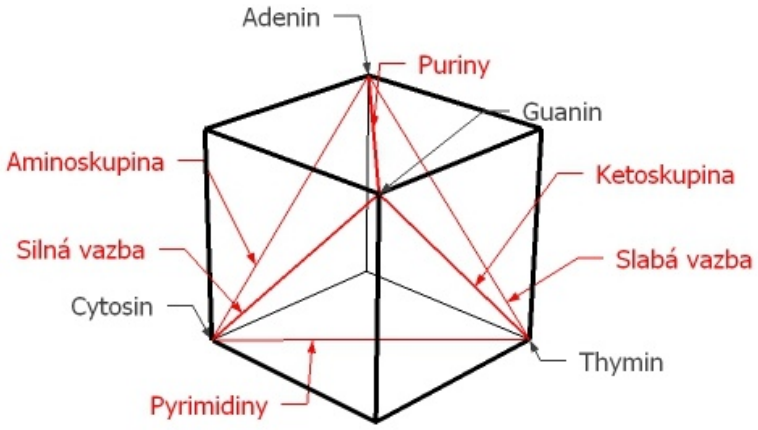


Obr. 3.3: RGB barvení zkráceného genomu (1000bp) *Chara vulgaris*

středu krychle do vrcholů mají jsou: [11, 1, 34]

$$\begin{aligned} \vec{a} &= \vec{i} + \vec{j} + \vec{k} \\ \vec{c} &= -\vec{i} + \vec{j} - \vec{k} \\ \vec{g} &= -\vec{i} - \vec{j} + \vec{k} \\ \vec{t} &= \vec{i} - \vec{j} - \vec{k} \end{aligned} \tag{3.5}$$

Výhodou této metody reprezentace genomických dat je, že zachovává vlastnosti bazí, jak je patrné z obrázku 3.4. Současně je výhodná pro hledání vzorových sekvencí a periodických složek sekvence s pomocí Fourierovy transformace. [1]



Obr. 3.4: Reprezentace krychlí

3.3.4 Reprezentace komplexním číslem

Reprezentace v komplexní rovině odráží taktéž komplementaritu a další chemické vlastnosti bazí, v tomto případě přiřazením komplexních čísel k nukleotidům. Jde

o dvourozměrné mapování, které získáme jednoduchým sklopení reprezentačního tetrahedronu do roviny. Kterou rovinu sklopíme záleží na požadovaných vlastnostech výsledného signálu. Sklopením přední strany krychle, získáme osy y a z, vypustíme osu x nesoucí informaci o druhu navázaného radikálu (amino-, keto- skupina) [11] Přiřazení komplexních souřadnic jednotlivým bazím je libovolné a reflektuje biologické skutečnosti (typ báze, druh vazby a volného radikálu). Například přiřazení 3.6

$$\begin{aligned} \vec{a} &= 1 + \jmath \\ \vec{c} &= -1 - \jmath \\ \vec{g} &= -1 + \jmath \\ \vec{t} &= 1 - \jmath \end{aligned} \tag{3.6}$$

vychází z požadavku na zachování komplementarity bazí, což je vyjádřeno tím, že komplementární dvojice jsou namapovány napříč a mají opačná znaménka. Dvojice A–T má kladnou reálnou část a naopak dvojice G–C má reálnou část zápornou.

$$\begin{aligned} \vec{a} &= -1 + \jmath \\ \vec{c} &= -1 - \jmath \\ \vec{g} &= 1 + \jmath \\ \vec{t} &= 1 - \jmath \end{aligned} \tag{3.7}$$

Přiřazení 3.7 má výhodu v tom, že dvojice komplementárních sekvencí DNA v signálové podobě má shodné absolutní hodnoty, lišící se znaménkem, takže jejich součet je vždy nulový. Pokud přiřadíme bazím pouze reálnou či imaginární část, získáme reprezentaci 3.8, která zachovává nulový součet komplementárních sekvencí a současně mají komplementární dvojice buď pouze reálnou nebo komplexní hodnotu. [11]

$$\begin{aligned} \vec{a} &= -1 \\ \vec{c} &= -\jmath \\ \vec{g} &= \jmath \\ \vec{t} &= 1 \end{aligned} \tag{3.8}$$

Fázová analýza vychází z komplexního mapování DNA sekvence. Argument nebo fáze komplexního čísla je úhel s kladnou reálnou osou. Při násobení celými násobky 2π se fáze nezmění, pro odstranění dvojznačnosti se obor hodnot stanovil na $(-\pi, pi]$ rad. [10, 13, 11]

Při použití mapování 3.6, které zachovává informace o síle vazby a typu báze, jsou jednotlivé fáze:

$$\begin{aligned} \vec{a} = 1 + \jmath \rightarrow \quad & \varphi_a = \frac{\pi}{4} \\ \vec{c} = -1 - \jmath \rightarrow \quad & \varphi_c = -\frac{3\pi}{4} \\ \vec{g} = -1 + \jmath \rightarrow \quad & \varphi_g = \frac{3\pi}{4} \\ \vec{t} = 1 - \jmath \rightarrow \quad & \varphi_t = -\frac{\pi}{4} \end{aligned} \tag{3.9}$$

3.3.5 Kumulovaná fáze

Jde o kumulativní součet všech fází komplexních čísel od počátku sekvence až po aktuální pozici. Vektor kumulativního součtu je pak zobrazen v grafu a analyzován. Dá se získat též z četností výskytu jednotlivých nukleotidů ze vzorce

$$s_c = \frac{\pi}{4}[3(G_n - C_n) + (A_n - Tn)], \tag{3.10}$$

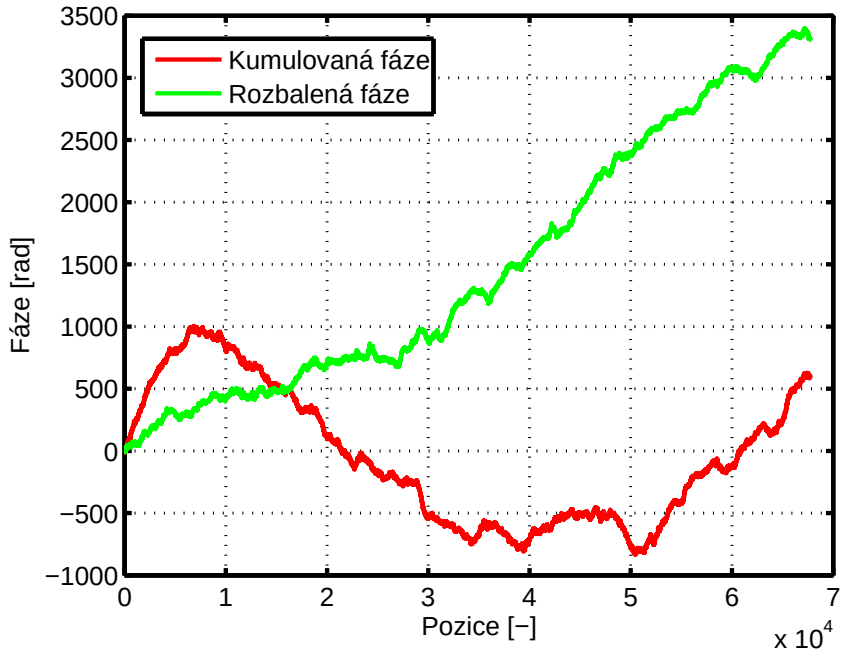
kde G_n, C_n, A_n, Tn jsou kumulativní četnosti bazí až po aktuální pozici. Sklon křivky kumulované fáze ukazuje na poměr výskytu jednotlivých bazí, jež se odvíjí od DNA konkrétního původu. [12, 38]

3.3.6 Rozbalená fáze

Rozbalená fáze je korigovanou fází elementů v sekvenci. Absolutní hodnota rozdílu fází dvou po sobě jdoucích je záměrně snižována na hodnotu menší než π přičítáním či odečítáním vhodného násobku čísla 2π. Tím dojde k eliminaci velkých rozdílů ve fázové sekvenci. Na rozdíl od kumulované fáze rozbalená fáze ukazuje na výskyty tranzicí mezi nukleotidy. Při reprezentaci 3.9 jsou kladné tranzice $A \rightarrow G, G \rightarrow C, C \rightarrow T, T \rightarrow A$, jež určují kladný přírůstek křivky. Negativní tranzice $A \rightarrow T, T \rightarrow C, C \rightarrow G, G \rightarrow A$ určují záporný přírůstek křivky. Ostatní tranzice jsou neutrální, nedochází při nich ke změně fáze. Sklon křivky rozbalené fáze je

$$\frac{2}{\pi} s_u = f_+ - f_-, \tag{3.11}$$

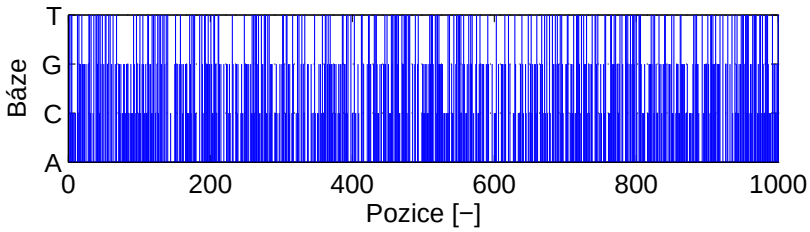
kde f_+ je frekvence kladných tranzic nukleotidů a f_- frekvence záporných tranzic.[10, 13]



Obr. 3.5: Fázova analýza mitochondriálního genomu *Chara vulgaris*

3.3.7 Reprezentace celým číslem

Jednorozměrná reprezentace nukleotidů číslicí z množiny $(0, 1, 2, 3)$ Je možných 24 různých přiřazení těchto číslic, avšak snahou je získat nejvíce monotonní reprezentaci [11]. To splňuje přiřazení $T = 0 \quad C = 1 \quad A = 2 \quad G = 3$ Toto přiřazení však zvýhodňuje puriny nad pyrimidiny. Je možné a vhodné zvolit přiřazení dle požadavků další analýzy dat. Nevýhodou libovolného přiřazení je zavedení nechtěných matematických vlastností (implikovaná nerovnost mezi bazemi). [1, 23]



Obr. 3.6: Reprezentace celým číslem zkráceného genomu (1000bp) *Chara vulgaris*

3.3.8 Reprezentace reálným číslem

Na rozdíl od předchozí reprezentace přiřazujeme nukleotidům reálná čísla. Nejčastější namapování je $A = -1,5 \quad T = 1,5 \quad C = 0,5 \quad G = -0,5$, které zohledňuje

komplementaritu bazí a je efektivní v hledání úseku vlákna DNA v sekvenci, nejčastěji pomocí AR (autoregresních) modelů. [1]

3.4 Fyzikálně-chemicky podmíněné mapování

Mapování nukleotidů při fyzikálně-chemické reprezentaci je založeno na skutečných vlastnostech bazí. Skupina zahrnuje EEIP reprezentaci, reprezentaci atomovým číslem a reprezentaci nukleotidových párů.

3.4.1 EEIP reprezentace (*Electron-Ion Interaction Potential*)

V podstatě jde o reprezentaci reálnými čísly, která však plynou z fyzikální podstaty. Nukleotidovým bazím jsou přiřazený kvazi-valenční čísla a jde tak o reprezentaci energie volných elektronů podél sekvence DNA [1, 23]. Výsledný vektor sekvence je tak sledem hodnot, přiřazených jednotlivým nukleotidům:

$$\begin{aligned} A = 0,1260 \quad C = 0,1340 \\ G = 0,0806 \quad T = 0,1335 \end{aligned} \tag{3.12}$$

3.4.2 Reprezentace atomovým číslem

V tomto případě jde o obdobu reprezentace celým číslem, bazím jsou přiřazovány hodnoty jejich skutečných atomových čísel:

$$\begin{aligned} A = 70 \quad C = 58 \\ G = 78 \quad T = 66 \end{aligned} \tag{3.13}$$

3.4.3 Reprezentace nukleotidových párů

Mapování je v tomto typu reprezentace založeno na spárování nukleotidů a společným mapováním této dvojice, případně na reprezentaci číslem 1 pro zvolený typ nukleotidu a reprezentace číslem -1 pro všechny ostatní. Podobně jako u komplexního mapování, i zde je na výběr z různých přiřazení, vyzdvihujících daný rys sekvence DNA.[1]

- Purin-Pyrimidinové (RY) pravidlo – hodnota 1 pro puriny (A, G), hodnota -1 pro pyrimidiny (C, T)
- $A\vec{A}$ pravidlo – pokud $n_i = A$ potom $u_i = 1$, pro ostatní případy $u_i = -1$
- $T\vec{T}$ pravidlo – pokud $n_i = T$ potom $u_i = 1$, pro ostatní případy $u_i = -1$
- $G\vec{G}$ pravidlo – pokud $n_i = G$ potom $u_i = 1$, pro ostatní případy $u_i = -1$
- $C\vec{C}$ pravidlo – pokud $n_i = C$ potom $u_i = 1$, pro ostatní případy $u_i = -1$
- SW pravidlo vodíkové vazby – $u_i = 1$ pro páry se silnou vazbou (G, C) a $u_i = -1$ pro páry se slabou vazbou (A, T)
- Hybridní (KM) pravidlo – $u_i = 1$ pro nukleotidy A nebo C, $u_i = -1$ pro nukleotidy T nebo G

Nejpoužívanější je RY pravidlo, ale možné jsou i další volby reprezentace na základě různých charakteristik. Hlavním využitím této reprezentace je predikce genových a exonových oblastí. Využívá se znalosti toho, že introny v genomu jsou bohatší na výskyt nukleotidů A a T zatímco exony obsahující větší množství C a G. [1]

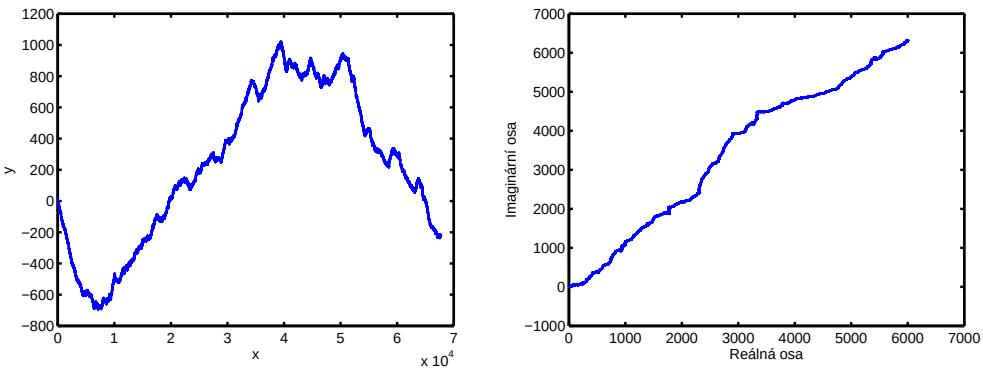
3.5 Grafické reprezentace DNA sekvencí

3.5.1 DNA walk

DNA walk je metoda pro zobrazení DNA sekvence a pozorování korelací v širokém rozsahu. Pro konvenční jednorozměrný náhodný DNA walk je zvolena reprezentace $[u(i) = +1]$ pro baze C a T nebo $[u(i) = -1]$ pro baze A a G. Celý proces si lze představit, jako chůzi při níž chodec provádí krok nahoru či dolů [33]. Lze tak snadno pozorovat trend poměru puriny-pyrimidiny, slouží tak opět k rozeznání exonových oblastí. [5, 1]

$$\begin{array}{llll} A & x(i) = 1 & G & x(i) = -1 \\ C & x(i) = \jmath & T & x(i) = -\jmath \end{array} \tag{3.14}$$

Pro hledání periodicit nebo nukleotidových struktur existuje dvojrozměrný DNA walk s komplexní reprezentací. DNA walk umožňuje graficky znázornit vývoj DNA sekvence a relativní trend výskytu nukleotidů v celé sekvenci. [5]



(a) (b)

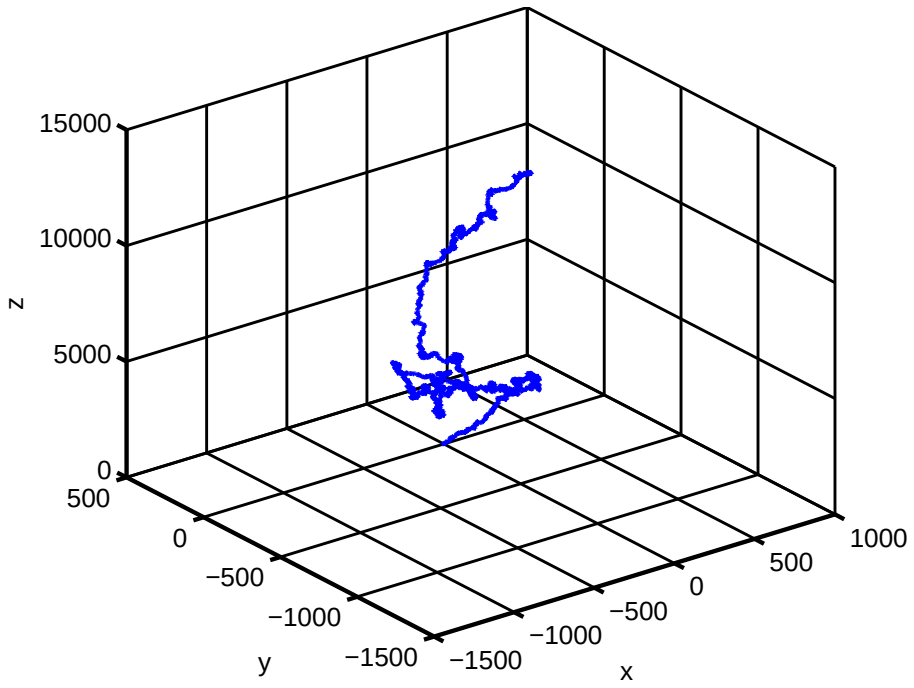
Obr. 3.7: DNA walk a) 1D b) 2D komplexní pro *Chara vulgaris*

3.5.2 Z křivka

Z křivka je křivka v trojrozměrném prostoru unikátní pro pro danou sekvenci DNA. Současně je možná zpětná rekonstrukce sekvence ze signálu, Z křivka nese celou informaci o DNA. Slouží jako grafická reprezentace sekvence, kdy je možné studovat jak globální tak lokální vlastnosti zmapované sekvence. Byla úspěšně použita k identifikaci replikačních počátků u několika DNA sekvencí organismů z říše Archea. Z křivka je složena z několika uzlů $P_0, P_1, ..., P_N$ se souřadnicemi x_n, y_n, z_n $(n = 0, 1, 2, ..., N)$, kde N je délka sekvence, které jsou dány Z-transformací 3.15 sekvence

$$
\begin{aligned}
x_n &= (A_n + G_n) - (C_n + T_n) \equiv R_n - Y_n \\
y_n &= (A_n + C_n) - (G_n + T_n) \equiv M_n - K_n \\
z_n &= (A_n + T_n) - (C_n + G_n) \equiv W_n - S_n \\
n &= 0, 1, 2, ..., N \quad x_n, y_n, z_n \in [-N, N],
\end{aligned} \tag{3.15}
$$

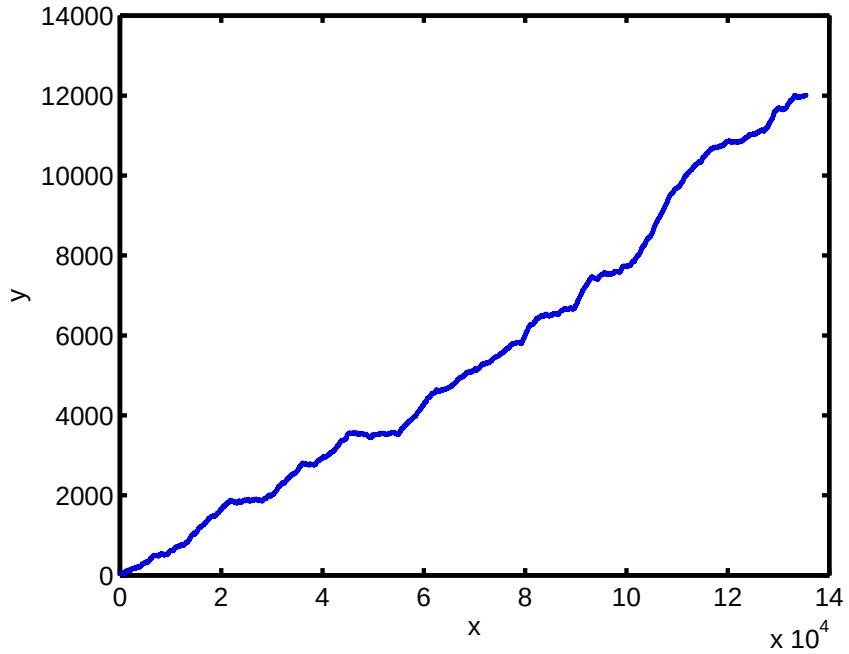
kde A_n, C_n, G_n, T_n jsou kumulované součty indikačních vektorů od první báze po bázi na pozici n v sekvenci. R, Y, M, K, WaS reprezentují vlastnosti nukleových kyselin typ báze(purinová, pyrimidinová), typ navázaného radikálu (amino, keto) a sílu vodíkové vazby (slabá, silná). Tyto vlastnosti jsou neseny odpovídajícími osami. Z křivka vždy začíná v počátku souřadné soustavy, neboť platí $A_0 = C_0 = G_0 = T_0 = 0$ a tedy $x_0 = y_0 = z_0 = 0$. [1, 45]



Obr. 3.8: Reprezentace Z křivkou mitochondriálního genomu *Chara vulgaris*

3.5.3 DV křivka

DV křivka je reprezentace bez degenerace a ztráty informace, taktéž zpětně rekonstruovatelná. Principem je přiřazení ne obvyklého jednoho vektoru, ale vektorů dvou po sobě jdoucích. Díky tomu se vyhne obvyklé degeneraci signálu tvorbou smyček, úseků které se kříží. Signálová sekvence se totiž podobně jako u jednorozměrného DNA walku vyvíjí pouze ve směru osy, jak je patrné z obrázků 3.9. Reprezentace



Obr. 3.9: Reprezentace DV křivkou mitochondriálního genomu *Chara vulgaris*

DV křivkou má několik výhodných vlastností. Je jednoduchá bez potřeby upravování dalších parametrů. Reprezentace bazí A a G je symetrická, taktéž reprezentace bazí T a C. A pokud je hodnota na konci grafu větší než nula, nachází se v sekvenci převaha adeninové báze, zatímco hodnota pod osou x značí převahu guaninové báze. Tato grafická reprezentace je výhodná především k zobrazení mutací v sekvencích, neboť každá bodová mutace je snadno viditelná. [46]

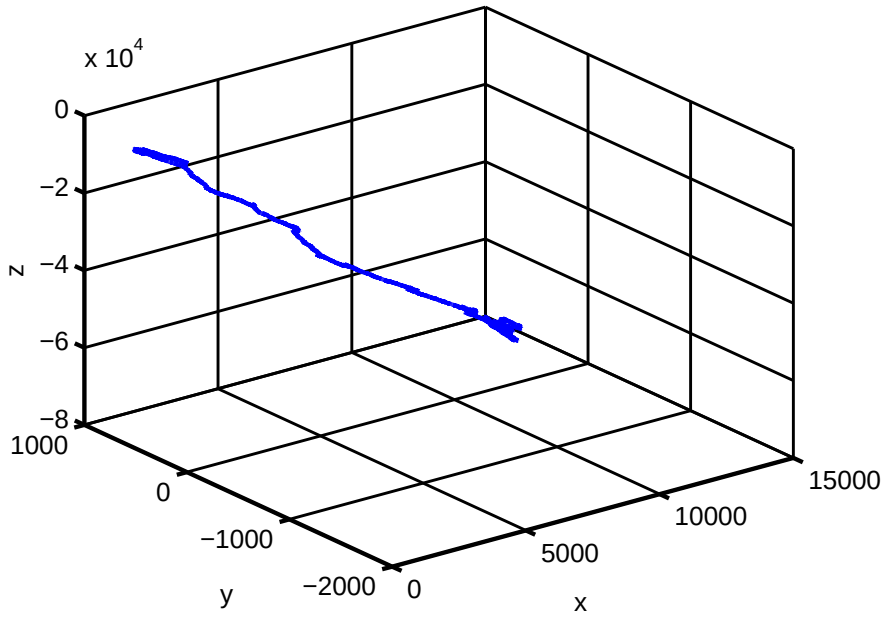
3.5.4 H křivka

V podstatě jde o reprezentaci velmi podobnou krychlové reprezentaci s tím rozdílem, že vektory reprezentující báze směřují z vrcholu krychle (namísto ze středu jako u krychlové reprezentace). Vektory jsou složené z jednotkových vektorů ve tvaru 3.16

$$\begin{aligned} \vec{a} &= \vec{i} + \vec{j} - \vec{k} \\ \vec{c} &= -\vec{i} - \vec{j} - \vec{k} \\ \vec{g} &= -\vec{i} + \vec{j} - \vec{k} \\ \vec{t} &= \vec{i} - \vec{j} - \vec{k} \end{aligned} \tag{3.16}$$

Výsledný tvar v prostoru neustále klesá díky zápornému jednotkovému vektoru ve směru osy z, hodnota konce odečtená na ose z také udává délku zmapované sekvence. Hodnota v ose y udává poměr purinů a pyrimidinů v sekvenci, hodnota na x-ové ose

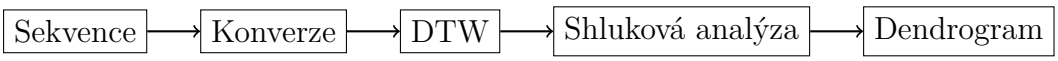
pak poměr síly vazeb. Podobně jako DV a Z křivka i zde jde o nedegenerativní mapování s možností zpětné rekonstrukce. [16, 36]



Obr. 3.10: Reprezentace H křivkou mitochondriálního genomu *Chara vulgaris*

4 KLASIFIKACE ORGANISMŮ

Dalším krokem po převedení znakové sekvence na číselnou sekvenci je samotná analýza a potřebné předzpracování. Celý algoritmus může být znázorněn na diagramu 4.1.



Obr. 4.1: Diagram algoritmu

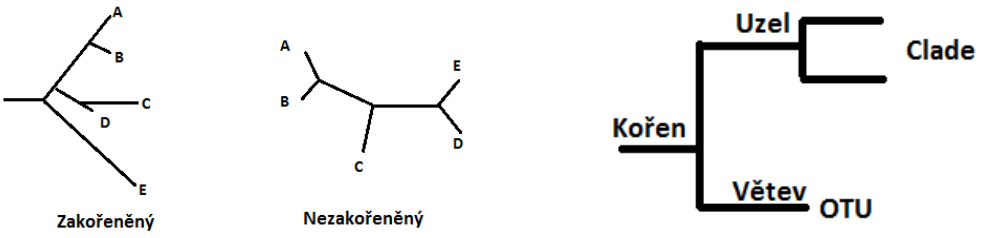
4.1 Fylogenetické stromy

Fylogenetický strom (či jinak dendrogram) je výchozí metodou studia příbuzenských vztahů mezi organismy v rámci evoluční teorie. Nesledujeme při ní přímo příbuzenské vztahy, ale odvozené vztahy založené na podobnosti. Genealogická příbuznost je tak vysvětlením sledované strukturní podobnosti [14].

Výchozím objektem sledování je biologický druh, tedy operační taxonomická jednotka (OTU), reprezentovaná genovou sekvencí. Operační taxonomické jednotky jsou zaneseny do grafu – dendrogramu. Ten spojuje OTU jednotlivými větvemi do uzlů, představující předpokládané společné předky spojených OTU. Všechny spojnice vycházející z uzlu, spolu s uzly na nich ležícími včetně OTU, tvoří klad (*clade*), větev vyššího řádu. Délka větví reprezentuje evoluční vzdálenost uzlů. Někdy délka větví nemá význam a zajímá nás pouze topologie stromu, jde o kladogram. [14, 7]

Stromy mohou být zakořeněné (*rooted*) či nezakořeněné (*unrooted*) v závislosti na umístění společného předka ve stromu. Pro dosažení zakořeněného stromu se využívá zařazení homologní, avšak vzdálené sekvence do dendrogramu, tzv. *outgroup*. Není-li k dispozici, musíme strom interpretovat jako nezakořeněný (viz.Obr 4.2). [14]

Metody vzniku dendrogramů založené na operačních jednotkách vyjádřených znakovou sekvencí DNA patří mezi znakové metody tvorby dendrogramů (*character state methods*). Principem je hledání konkrétního stromu z množiny možných dendrogramů, které je možno vytvořit. Určujícím parametrem může být například počet možných záměn, které vedly k vytvoření stromu a vybereme strom s nejmenším množstvím záměn (metoda maximální parsimonie). Metody založené na hodnotách podobnosti (vzdálenosti) mezi sekvencemi nazýváme vzdálenostní metody (*distance matrix methods*). Jde o tvorbu jednoho určitého stromu na základě vzdálenosti mezi jednotlivými sekvencemi, kdy OTU s nejmenší vzdáleností spojíme do uzlu (metoda UPGMA). [14, 7]



(a) (b)

Obr. 4.2: Rozdělení (a) a popis fylogenetických stromů (b)

Nejpoužívanější a v této bakalářské práci též využitou metodou tvorby dendrogramu je distanční metoda UPGMA (*Unweighted Pair Group Method with Arithmetic Mean*). UPGMA je hierarchická shlukovací metoda popsaná v roce 1977. Prvním krokem vypočtení vzdálenosti mezi sekvencemi. Pro znakové sekvence je používaná proporcionální vzdálenost (*p-distance*) ,popsaná jednoduchým vzorcem $p = n_p/n$, kde n_p je počet záměn (mutací) mezi dvěma sekvencemi a n je celkový počet pozic v sekvenci. Další možností výpočtu vzdálenosti je substituční model Jukes-Cantor vycházející z Markovových modelů. Ve vytvořené matici vzdáleností nalezneme nejnižší hodnotu, tedy nejmenší vzdálenost dvou sekvencí, a spojíme tyto v jeden uzel v dendrogramu. Přepočítáme matici vzdáleností, tak že spojenou dvojici sloučíme do jedné pozice a přiřadíme jim novou vzdálenost jako aritmetický průměr. Postup se opakuje do vytvoření kompletního dendrogramu. Metoda UPGMA nebere v potaz rozdílnou rychlost evoluce a předpoklad evoluce, jde o srovnání podobnosti fenotypů. [14]

Naopak metoda NJ již v potaz evoluční rychlost bere. Metoda NJ (*neighboor-joining*) vychází z konstrukce hvězdicovitého stromu. Postup je podobný metodě UPGMA. Metoda spojuje dvojice uzlů či OTU do větví s minimálním součtem větví. Spojená dvojice je v každém kroku nahrazena společným předkem, což sníží počet větví o jednu. Postup se opakuje až dokud nezbydou pouze bifurkace. Výstupem metody NJ je strom, který je i rekonstrukcí evolučního procesu, fylogram. [14]

Předpokladem pro správné vypočtení proporcionální vzdálenosti je zarovnání sekvencí. Kvalita dendrogramu je totiž podmíněna kvalitou přiřazení (zarovnání) sekvencí [14]. Cílem párového zarovnání sekvencí je ukázat na dostatečnou podobnost dvou sekvencí a rozhodnout tak, zda jsou tyto geny homologní. Podobnost je zde míněna jako kvantitativní určení této podobnosti, homologie genů je naopak závěrem, že tyto dvě sekvence sdílejí stejného evolučního předka. Geny buď jsou či nejsou homologní, neexistují zde stupně rozlišení jako u podobnosti. Mezi změny,

které se dějí během divergence ze společného předka patří substituce, inzerce a delece. V ideálním případě, kdy zarovnání odráží evoluční proces mezi dvěma sekvencemi, by rezidua (nezarovnané nukleotidy) reprezentovaly substituce. Místa, kde rezidua s ničím nekorespondují by reprezentovaly inzerce či delece. Tyto mezery jsou v zarovnání obvykle znázorněny znakem pomlčky. [4]

Zarovnání, které pokrývá v plné míře celou délku vstupní sekvence, nazýváme globální zarovnání. Mnoho proteinů však nevykazuje znaky podobnosti v rámci celé sekvence, ale naopak v rámci mozaiky modulárních domén. Globální zarovnání však tento jev není schopno zpracovat, neboť bylo zpracováno před objevem exonových a intronových úseků genů. Z toho vyplývá nutnost lokálního zarovnání. Takové zarovnání sestává z párů subsekvencí, které mohou být obklopeny nesouvisejícími rezidui.[4]

Dnes známý Needleman-Wunschův algoritmus je aplikace dynamického programování (převedení problému na hledání nejlepší trasy s daným kritériem optimality - skóre) na hledání optimálního zarovnání sekvencí. Metoda porovnává každý možný pár v sekvenci a generuje zarovnání krok za krokem. Výsledná trasa zarovnání směřuje od jednoho konce matice do opačného, generuje tedy globální zarovnání. Jednoduchou modifikací algoritmu můžeme dosáhnout lokálního zarovnání, jde o Smith-Watermanův algoritmus. Trasa optimálního zarovnání nemusí dosahovat okrajů, ale může skončit kdekoliv uvnitř matice. Takové zarovnání bude lokálně optimální, pokud jeho skóre nebude vhodnější při prodloužení či zkrácení rozsahu zarovnání. Jak Needleman-Wunschův tak Smith-Watermanův algoritmus závisí na volbě skórovacího systému, který vyjadřuje pravděpodobnost změny jednoho nukleotidu či aminokyseliny v jiný nukleotid nebo aminokyselinu na základě fyzikálně-chemických a evolučních měřítek. [4, 31]

4.2 Dynamické borcení časové osy

Metoda dynamického borcení časové osy (též. dynamic time warping, DTW) je metoda původně navržená pro rozpoznání mluveného slova a z ní plynoucí nutnosti zarovnání obou signálů v čase. Jde o algoritmus pro dva jednorozměrné signály řešený též dynamickým programováním. Délka obou signálů je algoritmem přizpůsobitelná a signály mohou být i zkráceny. Signály, jež vyjádříme jako sekvenci doplňkových vektorů:

$$\begin{aligned} A &= a_1, a_2, ..., a_i, ..., a_I \\ B &= b_1, b_2, ..., b_j, ..., b_J \end{aligned} \tag{4.1}$$

se vynesou podél os $i - j$ roviny. Rozdíly v časovém posunu můžeme znázornit jako funkci $F = c(1), c(2), ..., c(k), ..., c(K)$, kde bod $c(k) = (i(k), j(k))$. Dále stanovíme

vzdálenost mezi dvěma doplňkovými vektory $d(c) = d(i,j) = \|a_i - b_j\|$. Časově normalizovaná vzdálenost mezi signály A a B je definována následovně:

$$D(A,B) = \frac{1}{N} \times Min_F \left[\sum_{k=1}^{K} d(c(k)) \times w(k) \right] \tag{4.2}$$

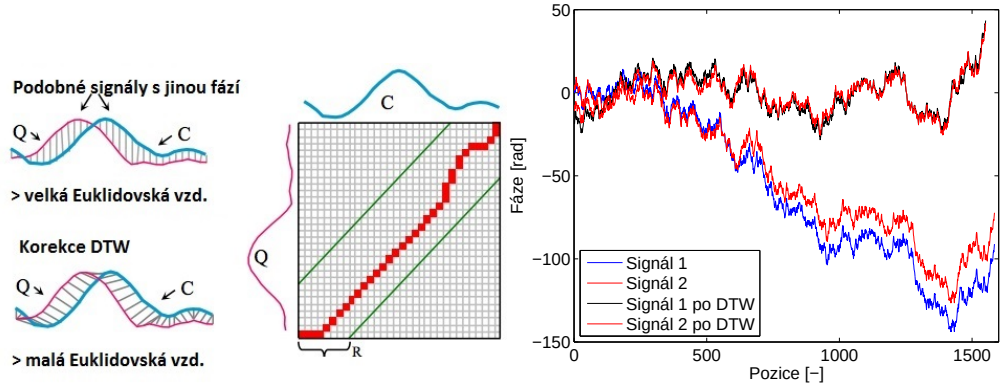
kde w(k) je nezáporný váhovací koeficient a výraz uvnitř je vahovaná kumulovaná vzdálenost funkce F. Řešení pomocí dynamického programování můžeme aplikovat na rovnici 4.4 a převést na algoritmus s počáteční podmínkou 4.3

$$g(1,1) = 2d(1,1) \tag{4.3}$$

Rovnice pro dynamické programování má tvar:

$$g(i,j) = Min \begin{cases} g(i,j-1) + d(i,j) \\ g(i-1,j-1) + 2d(i,j) \\ g(i-1,j) + d(i,j) \end{cases} \tag{4.4}$$

Matice kumulované vzdálenosti mezi signály se obvykle počítá z levého spodního rohu a celý algoritmus je tak podobný algoritmům pro zarovnání znakových sekvencí DNA v předchozí podkapitole. Samotné zarovnání je provedeno zpětným nalezením trasy matící.[37, 38] Aplikace a přínos metody DTW pro naši aplikaci je znázorněn na obrázku 4.3.



(a) (b)

Obr. 4.3: Znázornění metody DTW (a) a srovnání signálů před a po DTW (b)

Pravidlo omezení sklonu křivky F požaduje, aby pokud se bod c(k) posunul podél jedné z os m-krát, musí se posunout podle další osy n-krát než se znovu posune v původním směru. Toto může být zhodnoceno jako měřítko $P = n/m$. Pokud je $P = 0$ neplatí žádné omezení na sklon křivky, pokud $P = \infty$ pak je sklon křivky omezen na diagonální linii. Mezi další podmínky pro použití dynamického borcení

času patří podmínka monotónnosti, spojitosti pro zarovnávací funkci F. Z těchto omezení plyne nevhodnost algoritmu pro přílis odlišné signály co se počtu vzorků týče a průběhu týče. Dále je nutné s ohledem na jejich průběh zbavit vstupní signály lineárního trendu. [37]

4.3 Shluková analýza

Tvorba dendrogramů jak ze znakových sekvencí, tak z číselných sekvencí (signálů) stojí na pojmu shluková analýza. Shluková analýza je metoda pro seskupování objektů do skupin (shluků) tak, že objekty ve stejné skupině jsou si navzájem více podobné a současně méně podobné se zbylými objekty. Metody shlukové analýzy dělíme na hiararchické a nehierarchické. Výstupem nehierarchických shlukových metod jsou shluky bez rozlišení vazeb - hierarchie. Naopak výstupem hierarchických metod je dendrogram. Stěžejním tématem práce jsou dendrogramy, takže bude zmíněna pouze hierarchická metoda. [17, 21]

Mějme matici X rozměrů $n \times p$, kde n je počet objektů a p počet proměnných. Dále je nutné vypočítat vzdálenost mezi jednotlivými objekty v matici X. Volba metriky je pro roztřídění důležitá. Výchozí metrikou je Minkowského vzdálenost

$$d(x_i, x_j) = \sqrt[p]{\sum_{k=1}^{n} |(x_i k - x_j k)|^p}. \tag{4.5}$$

Pro $p = 1$ získáme vzdálenost Manhattan (city block), pro $p = 2$ získáme Euklidovu vzdálenost a pro $p = \infty$ získáme Čebyševovu vzdálenost. V práci je použita Euklidova vzdálenost:

$$d(x_i, x_j) = \sqrt{\sum_{k=1}^{n} (x_i k - x_j k)^2}. \tag{4.6}$$

Výpočtem vzdálenosti pro všechny objekty získáme matici podobností s prázdnou diagonálou. Dalším krokem je tvorba shluků na základě kritérií. Výchozí a v práci použitou metodou je UPGMA, zmíněná v kapitole o fylogenetických stromech. Metoda UPGMA upravuje matici podobností tak, že z objektů s nejmenší vzdáleností vytvoří shluk a přepočte matici podobností s použitím aritmetického průměru. Naopak metoda CLINK (complete linkage clustering) přepočte matici tak, že vezme pouze větší ze vzdáleností v matici podobností. Metoda SLINK (single linkage clustering) dosazuje naopak tu menší z dvojice vzdáleností. [17, 21]

4.4 Korelační koeficient

Matici podobností můžeme získat i z vytvořeného dendrogramu a srovnat s výchozí maticí podobností. Tak zjistíme míru zkreslení zavedení vytvořením dendrogramu.

K tomu slouží Pearsonův korelační koeficient. [21]

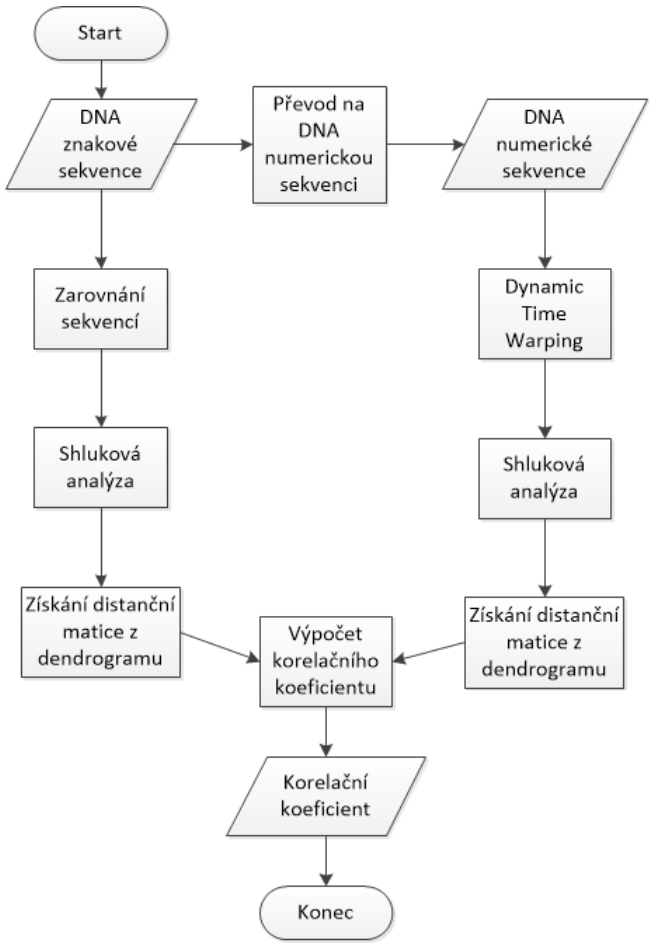
$$r_{xy} = \frac{\sum xy - \frac{1}{n}(\sum x)(\sum y)}{\sqrt{\left[\sum x^2 - \frac{1}{n}(\sum x)^2\right]\left[\sum y^2 - \frac{1}{n}(\sum y)^2\right]}} \tag{4.7}$$

$$-1 \leq r_{xy} \leq 1$$

Hodnota korelačního koeficientu nabývá hodnot $< -1, 1 >$. Hodnota koeficientu -1 značí zcela nepřémou závislost (antikorelaci), tedy nepřímou úměrnost. Hodnota 1 značí zcela přímou závislost. Při korelační koeficientu 0 není mezi znaky žádná lineární závislost (nelineární závislost být může). Hodnoty korelačního koeficintu nad $0,8$ se považují za dostačující k potvrzení korelace. [21]

5 PROGRAMOVÉ ŘEŠENÍ

Výše popsané teoretické poznatky byly aplikovány na tvorbu aplikace s grafickým uživatelským rozhraním (GUI) v programovacím rozhraní Matlab (verze 2012b/64b). Cílem této aplikace bylo umožnit konverzi DNA znakových sekvencí do numerické podoby s možnou další klasifikací. Výsledná podoba GUI je na obrázku 5.2. Grafické rozhraní umožňuje jednoduše provést analýzu potřebnou pro srovnání vhodnosti několika numerických reprezentací z hlediska jejich následné klasifikace. Celý algoritmus použitý pro zhodnocení využitelnosti numerických reprezentací, a který je i obsahem aplikace, je znázorněn na vývojovém diagramu 5.1.



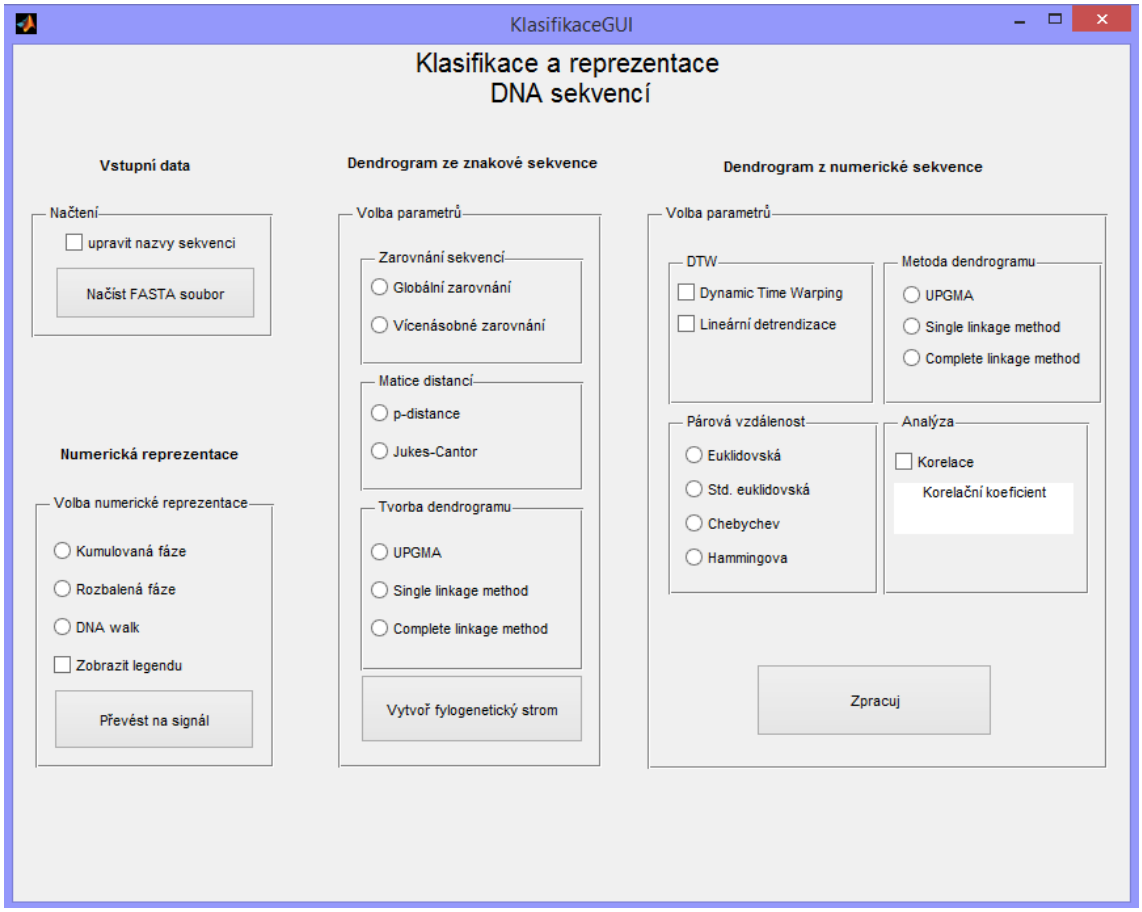
Obr. 5.1: Vývojový diagram aplikace

Vstupním krokem je načtení sekvenci ve datovém formátu FASTA obsahující vícečetné sekvence. Volitelnou možností před načtením sekvencí je zkrácení názvů sekvencí o popisné informace zobrazovaných ve výstupech programu.

Dalším krokem je převod na numerickou sekvenci podle uživatelem zvolené reprezentace. Byly vybrány tři numerické jednorozměrné reprezentace (kumulovaná

a rozbalená fáze, DNA walk), z důvodu možnosti aplikace zarovnání dynamickým borcením časem (určeným pro 1D signály), i když už byl popsán i algoritmus DTW pro vícerozměrné signály [43]. Spuštěním převodu stisknutím tlačítka „Převést“ se sekvence převedou na numerické sekvence a zobrazí se jejich grafický průběh. Volitelnou možností je zobrazení či skrytí legendy grafu.

Následuje volba parametrů pro tvorbu dendrogramu ze znakových sekvencí. Kvalita dendrogramu závisí na provedeném zarovnání. Zarovnat sekvence lze globálním párovým zarovnáním nebo vícenásobným zarovnáním. Ze zarovnaných sekvencí se vypočte distanční matice s využitím proporcionální vzdálenosti mezi dvojicemi sekvencí nebo evolučního modelu Jukes-Cantor. Pro shlukování lze vybrat jednu ze shlukovacích metod popsaných v kapitole 4.3: UPGMA, SLINK, CLINK. Výstupem bloku je dendrogram považovaný za referenční a též z dendrogramu získaná distanční matice pro výpočet korelačního koeficientu v dalším bloku aplikace.



Obr. 5.2: Grafické uživatelské rozhraní aplikace

Dalším blokem aplikace je tvorba dendrogramu z numerických sekvencí. Sekvence lze zarovnat metodou DTW s možností lineární detrendizace (proložení lineární křivkou a odečtení od signálu). V případě nezarovnávání metodou DTW jsou

numerické signály oříznuty pro potřeby shlukové analýzy. Dalším krokem je volba metriky pro vypočtení distanční matice, popsané v kapitole 4.3. Lze zvolit z Euklidovské vzdálenosti, standardizované Euklidovské vzdálenosti, Čebyševovy vzdálenosti a Hammingovy vzdálenosti. Shluková analýza probíhá stejným způsobem jako v předchozím bloku, na výběr jsou metody UPGMA, SLINK, CLINK. Výstupem je opět dendrogram.

Dodatečnou možností je vypočtení korelačního koeficientu mezi dendrogramem ze znakových sekvencí a z numerických sekvencí. Při hodnotě korelačního keoficientu nad $0,8$ se považuje výsledek za vyhovující, tedy oba stromy jsou si podobné.

Výhoda aplikace je v blokovém schématu práce, uživatel tak sice musí aplikaci v průběhu práce ovládat, ale výhodou je nezávislost bloků. Chce-li např. uživatel vytvořit pouze fylogenetický znakový strom postačí mu načíst data a provést výpočet dendrogramu. Chce-li naopak vytvořit dendrogram z numerické reprezentace, nemusí provádět tvorbu stromu ze znakových sekvencí. Tvorba znakového stromu je nutná v případě, že uživatel chce získat korelační koeficient mezi dendrogramy. V případě, že potřebuje korelační koeficienty z různých kombinací nastavení, zobrazuje se vždy aktuální vypočtený (v případě, že je checkbox Korelace aktivovaný). Lze tak zjistit korelační koeficient bez zpětného znovuvytváření stromů. Výstup oznamovacího okénka pro korelační koeficient slouží také jako indikátor chyb v postupu.

6 KLASIFIKACE ORGANISMŮ

V následující kapitole budou prezentovány výsledky zhodnocení využitelnosti DNA signálů (numerických sekvencí DNA) pro tvorbu dendrogramů. Cílem analýzy je srovnat znakové dendrogramy (dendrogramy z klasické znakové sekvence DNA) s dendrogramy z numerické sekvence DNA (získané po převodu numerickou reprezentací). Pro analýzu byly vybrány tři numerické reprezentace: kumulovaná fáze, rozbalená fáze a reprezentace DNA walk, popsané v kapitolách 3.3.4 a 3.5.1. Tyto reprezentace byly vybrány z důvodu jejich rozšířenosti, vývojem podél časové osy a výpočetní nenáročnosti. Vzhledem k návrhu analýzy spojené s aplikací dynamického borcení časové osy je právě jejich jednorozměrnost hlavním kritériem.

Analýza bude zaměřena na mitochondriální sekvence organismů z rostlinné říše (*Plantae*). Jde o velikostně i obsahově velmi rozmanitou skupinu sekvencí, rozdíly počtu nukleotidových bazí mezi jednotlivými mitochondriálními genomy jsou značné (200kbp - 2 Mbp) [26]. Z tohoto důvodu nejsou vhodné pro celogenomovou klasifikaci ani pro aplikaci DTW (viz. velikosti genomů a údaje z NCBI obsaženy v příloze A). Pro provedení genomové klasifikace by musely být sekvence převedeny na stejnou velikost (komprimací, podvzorkováním nebo zkrácením). Z tohoto důvodu se v práci pro klasifikaci používají genové úseky.

6.1 Taxonomie vybraných organismů

Jako výchozí genetický materiál pro analýzu bylo zvoleno 20 organismů z databáze NCBI, jejichž mitochondrie byly kompletně osekvenovány. Jde o organismy zařazené do rostlinné říše *Plantae* (viz. taxonomická klasifikace 6.1), převážně jde o hospodářské či jinak ekonomicky přínosné plodiny. Celkový počet všech kompletně osekvenovaných rostlinných mitochondriálních genomů je dle NCBI 114.

Z mitochondriálních genů byly pro klasifikaci zvoleny geny cox1, nad1 a nad4. Gen cox1 (*Cytochrome c oxidase subunit I*) je gen kódující hlavní podjednotku membránového transportního proteinu cytochrom c oxidázy, který je poslední částí respiračního elektronového transportního řetězce. Má i prioritní význam při DNA barcodingu. DNA barcoding je metoda pro identifikaci druhů založená na krátkých úsecích sekvence DNA, které tak reprezentují celý genom [19, 35].Gen nad1 kóduje podjednotku 1 NADH dehydrogenázového komplexu. Často podléhá trans-splicingu, kdy jednotlivé úseky genu jsou roztroušeny v rámci celého genomu. [8]. Gen nad4 kóduje podjednotku 4 stejného komplexu jako nad1. Komplex 1 NADH je složen celkově z 26 podjednotek kódovaných jak jadernými, tak mitochondriálními geny. Jde o značně konzervovaný gen, jeho podobnost v rámci rostlinné říše dosahuje 90% [27].

Tab. 6.1: Taxonomické zařazení organismů

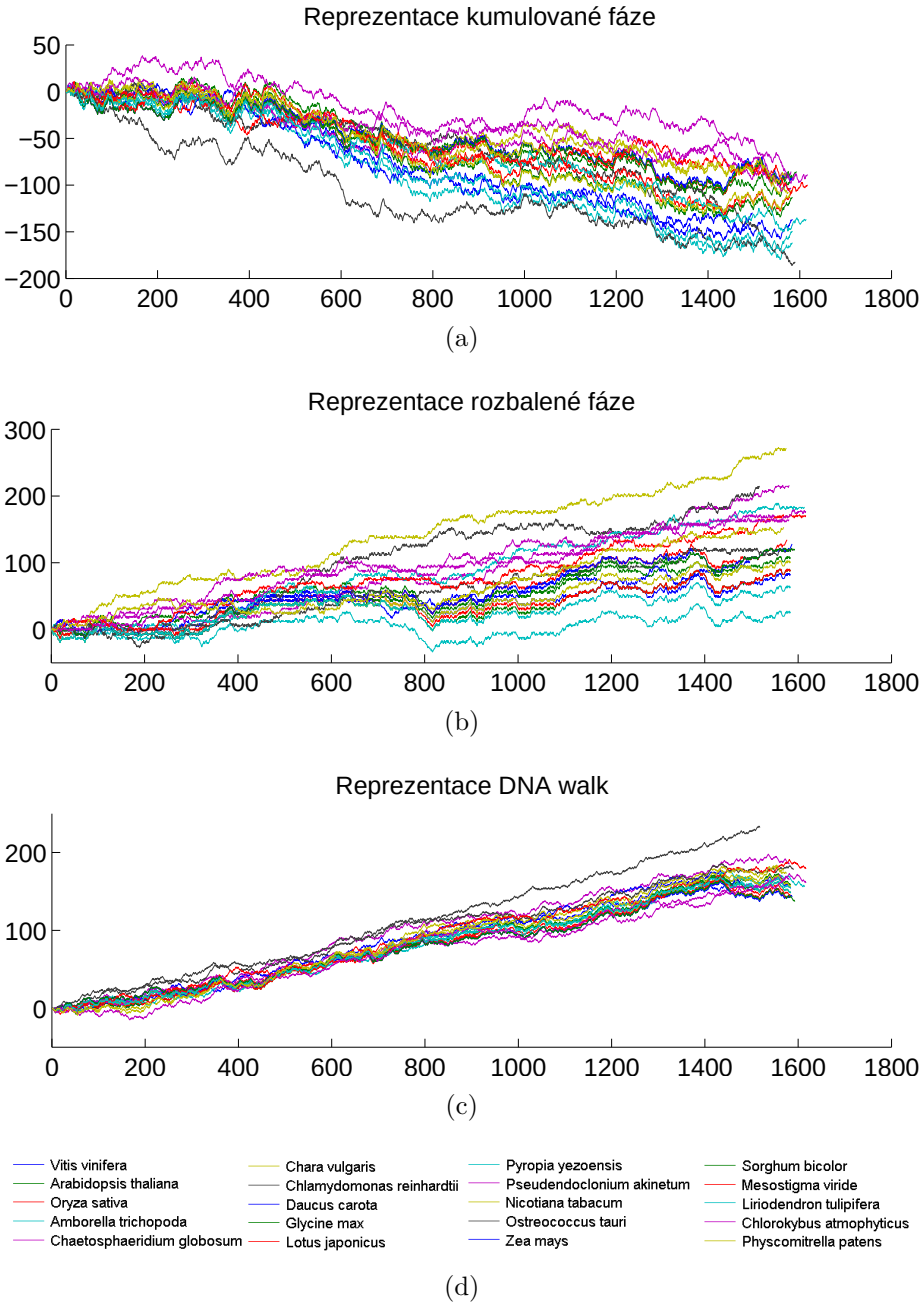
Latinský název	**Podříše**	**Třída**
Amborella trichopoda	Tracheophyta	Magnoliopsida
Arabidopsis thaliana	Tracheophyta	Rosopsida
Daucus carota	Tracheophyta	Rosopsida
Glycine max	Tracheophyta	Rosopsida
Chaetosphaeridium globosum	Charophyta	Coleochaetophyceae
Chara vulgaris	Charophyta	Charophyceae
Chlamydomonas reinhardtii	Chlorophyta	Chlorophyceae
Chlorokybus atmophyticus	Charophyta	Chlorokybophyceae
Liriodendron tulipifera	Tracheophyta	Magnoliopsida
Lotus japonicus	Tracheophyta	Magnoliopsida
Mesostigma viride	Charophyta	Mesostigmatophyceae
Nicotiana tabacum	Tracheophyta	Rosopsida
Oryza sativa	Tracheophyta	Liliopsida
Ostreococcus tauri	Chlorophyta	Mamiellophyceae
Physcomitrella patens	Bryophyta	Bryopsida
Pseudendoclonium akinetum	Chlorophyta	Ulvophyceae
Pyropia yezoensis	Rhodophyta	Bangiophyceae
Sorghum bicolor	Tracheophyta	Liliopsida
Vitis vinifera	Tracheophyta	Rosopsida
Zea mays	Tracheophyta	Liliopsida

6.2 Grafická reprezentace sekvencí

Všechny mitochondriální genomy byly stáhnuty z databáze NCBI a zkráceny na genové úseky. Znakové sekvence byly předeny dle zvolené numerické reprezentace na numerické sekvence s nimiž se dále pracovalo. Grafické průběhy numerických sekvencí pro úseky genu cox1 jsou níže na obrázku 6.1. Grafické průběhy zbylých genů jsou v uvedeny v příloze B.

Z grafických průběhů vidíme již první odlišnosti, výstupy reprezentace DNA walk jsou velmi podobné, naopak průběhy sekvencí rozbalené fáze a kumulované fáze jsou odlišnější. Euklidovská vzdálenost mezi sekvencemi fázové analýzy se zmenší po detrendizaci a zarovnání DTW. V příloze B máme grafické průběhy dalších genů nad1 a nad4. Pro průběhy sekvencí nad1 platí, že jsou si zde rozbalená fáze a DNA walk podobné, liší se výrazněji průběh kumulované fáze. U genu nad4 je dominantním

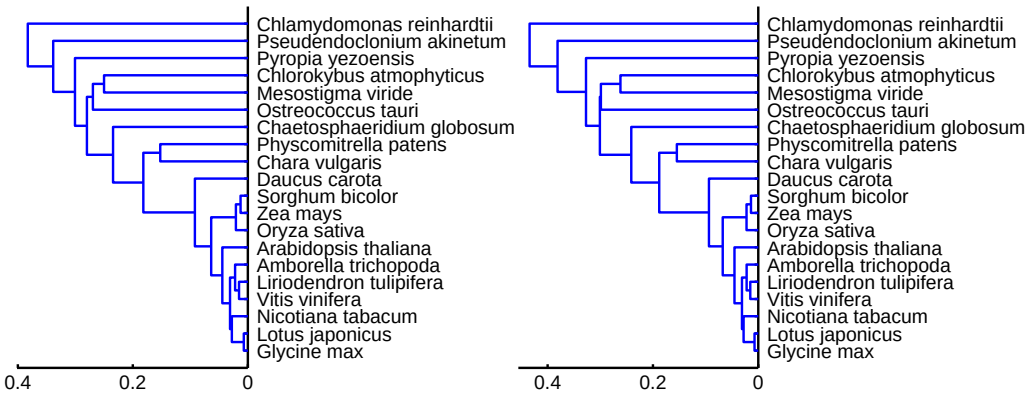
prvkem výrazně delší sekvence genu nad4 u organismu *Liriodendron tulipifera*. Zajímavé jsou i jakoby „opačné“ průběhy kumulované fáze a DNA walku pro úsek genu tohoto organismu.



Obr. 6.1: Průběhy numerických sekvencí DNA

6.3 Dendrogramy ze znakových sekvencí

Za referenční fylogenetický strom je v této práci považován strom získaný klasickými znakovými metodami. S takto vypočteným fylogenetickým stromem budou srovnávány ostatní stromy. Znakové sekvence byly zarovnány algoritmem párového globálního zarovnání i vícenásobným zarovnáním. Ze zarovnaných sekvencí byla vypočtena distanční matice proporcionálních vzdáleností a vytvořen dendrogram metodou UPGMA (viz obr.6.2). Dendrogramy vytvořené ze zbylých genových úseků jsou uvedeny v příloze B.

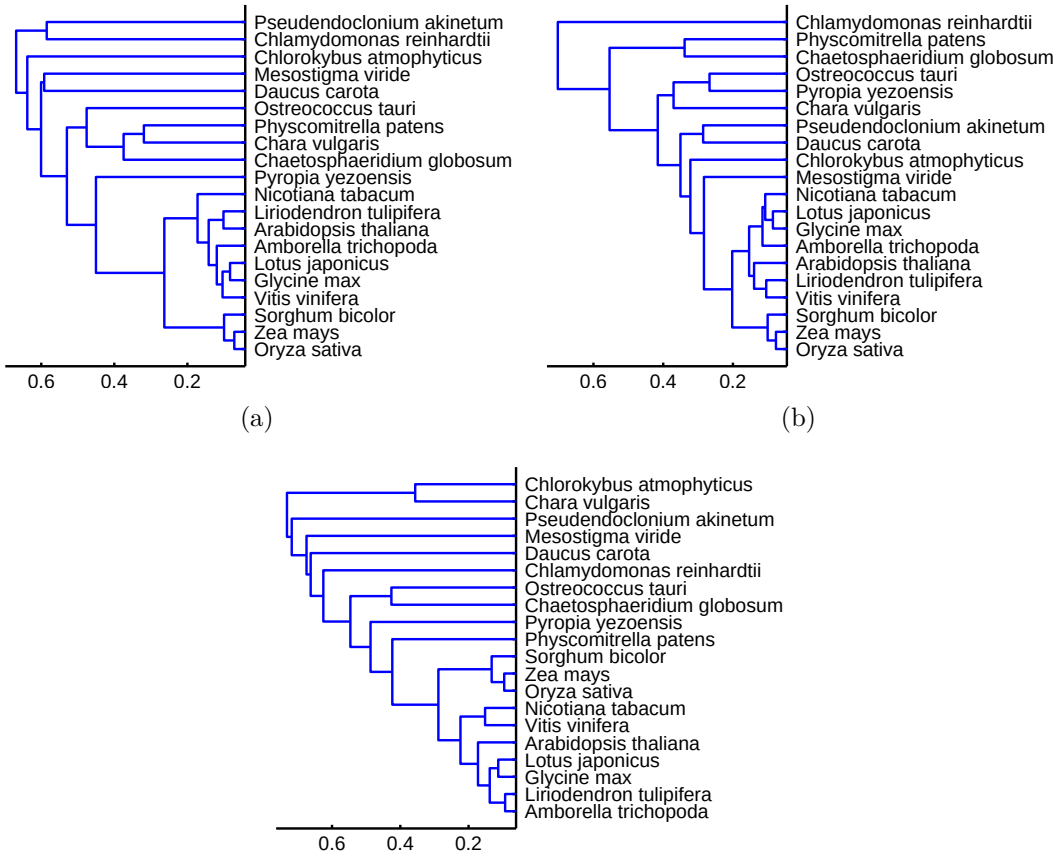


Obr. 6.2: Dendrogramy získané znakovými metodami globálním zarovnáním(a) vícenásobným zarovnáním(b) -gen cox1

Oba dendrogramy získáné z jinak zarovnaných sekvencí jsou co se shluků týče identické, liší se však vzdálenostmi. Vzhledem k nižší výpočetní náročnosti globálního zarovnání je upřednostněn v další analýze tento.

6.4 Dendrogramy z numerických reprezentací

Ze získaných numerických sekvencí, získaných na základě převodu na kumulovanou fázi, rozbalenou fázi a DNA walk, byly sestrojeny opět dendrogramy, které budou porovnány s dendrogramy ze znakových sekvencí. Sekvence organismů byly párově zarovnány metodou dynamického borcení času, byla mezi nimi vypočtena Euklidovská vzdálenost a sestrojen dendrogram metodou UPGMA (na obr. 6.3). Jednotlivé dendrogramy mají podobně rozvětvené větve, ale obsahem shluků se liší. Rozdíly mezi dendrogramy budou kvantitativně vyčísleny na základe korelačního koeficientu.



Obr. 6.3: Dendrogramy a)kumulované fáze b)rozbalené f. c)DNA walk - gen cox1

6.5 Výsledky srovnání dendrogramů

Stěžejním výstupem analýzy je výpočet korelačních koeficientů mezi jednotlivými dendrogramy. Srovnán byl vždy znakový strom, jak z globálně tak z vícenásobně zarovnaných sekvencí, se stromem získaným na základě numerických reprezentací. Mezi oběma stromy byl vypočten korelační koeficient srovnávající vzorky distančních matic obou dendrogramů. V tabulce 6.2 jsou uvedeny vypočtené korelační koeficienty mezi dendrogramy ze sekvencí genu cox1 v závislosti na volbě dalších parametrů.

Z tabulky lze pozorovat na přínos zarovnání numerických sekvencí metodou DTW, i přesto, že přínos DTW je u reprezentací rozbalené fáze a DNA walk nižší než u reprezentace kumulovanou fází. Podobně rozdíl koeficientů mezi globálním a vícenásobným zarovnáním je zanedbatelný. Na základě těchto výsledků bude DTW v analýze dalších genových sekvencí využit vždy a budou srovnávány jen dendrogramy z globálně zarovnaných sekvencí.Současně nelze na základě výsledků pro genové sekvence cox1 určit, která numerická reprezentace poskytuje nejlepší výsledky pro klasifikaci. Za povšimnutí stojí propad u korelačních koeficientů kumulované fáze

Tab. 6.2: Korelační koeficienty dendrogramů – cox1

Reprezentace	**Zarovnání znak.stromu**	**Bez DTW**	**s DTW**
Kumulovaná f.	Globalní	0,051	0,845
Kumulovaná f.	Vícenásobné	0,047	0,835
Rozbalená f.	Globalní	0,408	0,731
Rozbalená f.	Vícenásobné	0,389	0,723
DNA walk	Globalní	0,640	0,757
DNA walk	Vícenásobné	0,651	0,742

u nezarovnaných sekvencí (DTW).

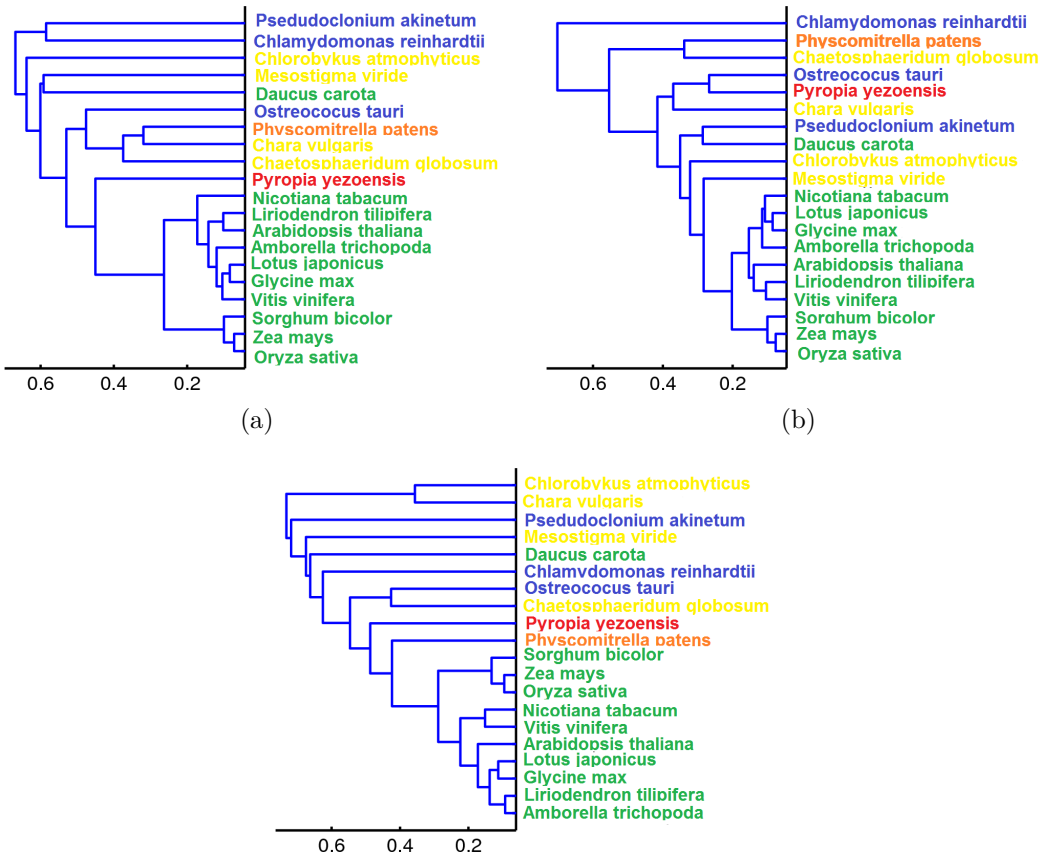
V druhé tabulce 6.3 jsou srovnány dendrogramy ze sekvencí všech genů cox1, nad1 a nad4, vždy globálně zarovnané (pro znakové sekvence) či zarovnané dynamickým borcením času (pro numerické).

Tab. 6.3: Korelační koeficienty pro různé reprezentace a genové úseky

Reprezentace	**cox1**	**nad1**	**nad4**
Kumulovaná f.	0,845	0,803	0,844
Rozbalená f.	0,731	0,848	0,860
DNA walk	0,757	0,858	0,877

Z daných výsledků je rozdíl mezi jednotlivými reprezentacemi malý, kumulovaná fáze ve všech případech dosáhla na hranici požadovaného korelačního koeficientu $0,8$ naopak reprezentace DNA walk dosáhla dvakrát nejvyššího korelačního koeficientu. Rozbalená fáze poskytla též přijatelné výsledky. Na základě těchto výsledků tak nelze jednoznačně určit nejvhodnější reprezentaci pro klasifikaci mitochondriálních sekvencí.

Dalším ze srovnání může být klasifikace dle taxonu Podříše. Na dendrogramech 6.4 jsou stejnou barvou vyznačeny organismy patřící do stejného taxonu. V ideálním případě by měli být zařazeny do stejných shluků, úseky genů v organismech však podléhají rozdílným evolučním vlivům a mohou být klasifikovány odlišně. Podíváme-li se na přiřazení organismů do shluků na základě příslušnosti k taxonu podříše, ani v tomto případě nelze považovat kvalitu jedné z reprezentací oproti ostatním z hlediska klasifikace za lepší.



Obr. 6.4: Dendrogramy z numerických sekvencí - taxonomická klasifikace a)kumulované fáze b)rozbalené f. c)DNA walk - gen cox1

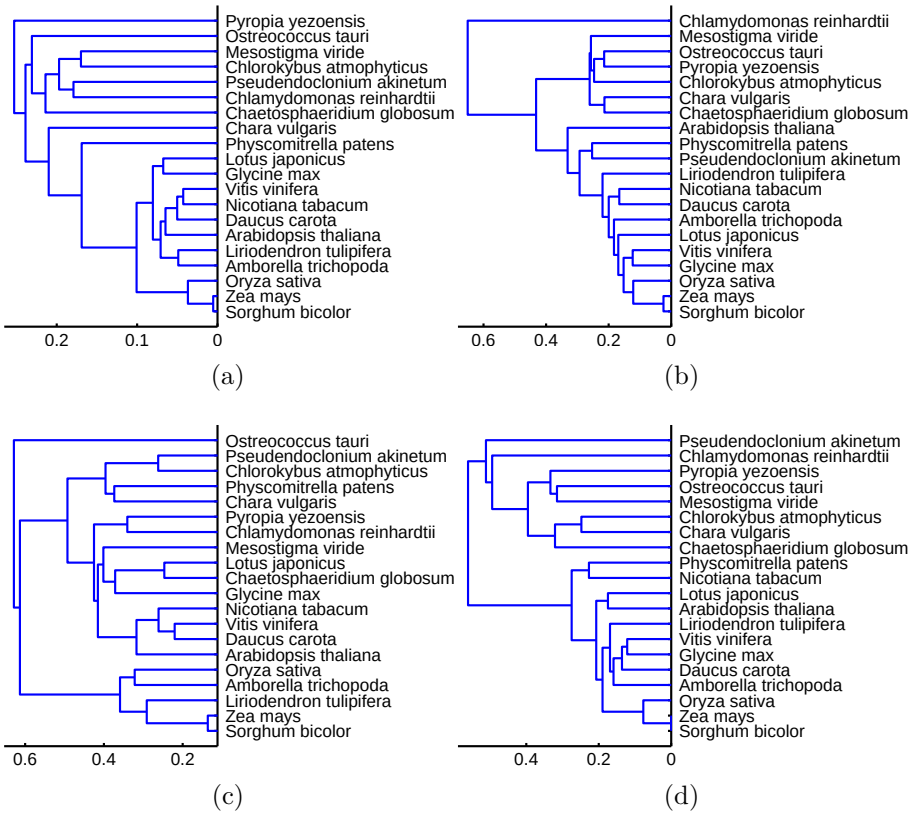
6.6 Srovnání s chloroplastovými sekvencemi

Chloroplastové sekvence spolu s mitochondriálními sekvencemi tvoří rostlinný mimojaderný genom. Z tohoto hlediska může být zajímavé analyzovat vhodnost chloroplastových sekvencí pro klasifikaci. Pro chloroplastové sekvence byly vybrány geny psaA, psbA a psbC. Gen psaA je gen kódující protein PSI P700 apoprotein A1, primárního elektronového donora v procesu fofosyntézy [20]. Gen psbA je podobně jako cox1 používán k DNA barcodingu [25] a spolu s psbC patří do skupiny genů kódujících podjednotky buněčného fotosystému I [22]. Analýza zde byla provedena podle stejného návrhu jako analýza mitochondriálních sekvencí. Analyzován byl i vliv zarovnání znakových sekvencí na korelační koeficient, naopak vliv DTW byl již vynechán. Výsledky pro gen psaA jsou v tabulce 6.4. Níže na obrázku 6.5 jsou dendrogramy jak pro výchozí znakový strom získaný z globálně zarovnaných sekvencí na základě proporcionální vzdálenosti a shlukovací metody UPGMA, tak dendrogramy z numerických sekvencí vypočtený z Euklidovských vzdáleností a sestrojen

Tab. 6.4: Korelační koeficienty dendrogramů – psaA

Reprezentace	**Zarovnání znak.stromu**	**Korelační koeficient**
Kumulovaná f.	Globalní	0,679
Kumulovaná f.	Vícenásobné	0,678
Rozbalená f.	Globalní	0,186
Rozbalená f.	Vícenásobné	0,183
DNA walk	Globalní	0,904
DNA walk	Vícenásobné	0,904

opět metodou UPGMA. Zbylé dendrogramy jsou v příloze B.



Obr. 6.5: Dendrogramy a)Znakové metody b)Kumulované f. c)Rozbalené f. d)DNA walk - gen psaA

V další tabulce 6.5 jsou získané korelační koeficienty ze všech genů a reprezentací. Jako referenční strom byl použit z globálně zarovnanách sekvencí UPGMA dendrogram.

Tab. 6.5: Korelační koeficienty genů psaA,psbA,psbC

Reprezentace	**psaA**	**psbA**	**psbC**
Kumulovaná f.	0,679	0,489	0,516
Rozbalená f.	0,186	0,436	0,119
DNA walk	0,904	0,631	0,618

Z této části analýzy vyplývá velký rozdíl mezi mitochondriálními a chloroplastovými sekvencemi, co se volby reprezentace ke klasifikaci týče. Narozdíl od mitochondriálních dendrogramů, kde nebylo možné určit nejvhodnější reprezentaci, můžeme zde hned vyloučit reprezentaci rozbalené fáze jako nejméně vhodnou. Nejlepšího výsledku zde dosáhla reprezentace DNA walk, i když na hranici $0,8$ dosáhla jen v případě genu psaA.

6.7 Shrnutí analýzy

Z analýzy mitochondriálních sekvencí a numerických reprezentací vyplynulo několik skutečností. Prvním z nich byl minimální rozdíl mezi globálně zarovnanými a vícenásobně zarovnanými sekvencemi při klasifikaci. Další skutečností byla nemožnost určit nejvhodnější reprezentaci pro klasifikaci mitochondriálních genů. Stabilní výsledky podala reprezentace kumulovanou fází, nejlepší však reprezentace DNA walk.

Z dodatkové analýzy chloroplastových sekvencí naopak jasně vyplynula vhodnost reprezentace DNA walk pro klasifikaci těchto sekvencí. Reprezentace rozbalenou fází se prokázala jako nejméně vhodnou pro klasifikaci chloroplastových sekvencí.

7 ZÁVĚR

Cílem práce bylo zhodnotit využitelnost numerických reprezentací na skutečných sekvencích pro klasifikaci organismů. Dílčími tématy tak byl popis biologických aspektů a výběr vhodných metod konverze DNA sekvence. V první části práce jsem popsal biologickou podstatu zadaného tématu, což bylo důležité téma. Bez porozumění rozdílům mezi jadernou a mitochondriální DNA a rozdílům mezi mitochondriální DNA živočichů a rostlin by bylo těžší se orientovat v dalších částech práce. Stručně lze shrnout, že hlavním rozdílem mezi jadernou a mitochondriální DNA je prokaryotní původ DNA u mitochondrií, který má cirkulární strukturu. Rozdílů mezi živočichy a rostlinami je více: živočišná mitochondriální DNA má odlišný genetický kód (rostlinná ne), neobsahuje nekódující úseky intronů (rostlinná ano) a mají odlišné uspořádání. Rostlinný mitochondriální genom je značně větší a dědí se maternálně.

V práci byl uveden dostatečný přehled metod konverze DNA do numerické podoby a byly uvedeny praktické ukázky na reálných sekvencích. Za zmínění stojí, že mnozí autoři nazývají stejné reprezentace odlišně a liší se i jejich setřídění. Z těchto různých metod byly vybrány tři pro následnou analýzu užitečnosti klasifikace organismů těmito metodami. Byla vybrána reprezentace DNA walk, kumulovaná a rozbalená fáze z důvodu jejich jednorozměrnosti, která se uplatní při shlukové analýze a zarovnání dynamickým borcením času.

Teoretické poznatky o fylogenetických stromech a jejich tvorbě spolu s poznatky o shlukové analýze, algoritmu dynamického borcení času a korelačním koeficientu jakožto základech pro klasifikaci organismů jsou součástí poslední teoretické části.

Praktickou části bylo vytvoření aplikace s grafickým uživatelským rozhraním v prostředí Matlab, která slouží ke klasifikaci organismů na základě numerických sekvencí. Aplikace nazvaná „KlasifikaceGUI“ umožňuje též kvantitativně porovnat dendrogramy na základě korelačního koeficientu. S využitím této aplikace byla provedena analýza a zhodnocení s cílem určit vhodnou numerickou reprezentaci pro klasifikaci organismů.

V poslední části práce bylo provedeno srovnání dendrogramů sestrojených na základě numerických sekvencí s dendrogramy sestrojenými klasickými znakovými metodami na vzorku 20 mitochondriálních genových úseků. Základním principem byl výpočet korelačního koeficientu mezi distančními matice získanými z obou dendrogramů, přičemž dendrogram ze znakových sekvencí byl referenční.

Výsledkem pro mitochondriální sekvence je malý rozdíl mezi reprezentacemi ve vhodnosti pro klasifikaci. Stabilní výsledky podala reprezentace kumulovanou fází, nejlepší výsledek však reprezentace DNA walk. Naopak pro dodatečně porovnávané chloroplastové sekvence vyplynula reprezentace DNA walk jako nejvhodnější, nao-

pak reprezentace rozbalenou fází se ukázala jako nevhodná pro klasifikaci. Srovnání reprezentací z hlediska klasifikace na vzorku mitochondriálních a chloroplastových sekvencí nebylo dosud publikováno a z analýzy vyplývá odlišná vhodnost reprezentací pro oba tyto typy mimojaderné DNA.

LITERATURA

[1] Abo-Zahhad, M.; Ahmed, S. M.; Abd-Elrahman, S. a.: Genomic Analysis and Classification of Exon and Intron Sequences Using DNA Numerical Mapping Techniques. *International Journal of Information Technology and Computer Science*, ročník 4, č. 8, 2012: s. 22–36, ISSN 20749007, doi:10.5815/ijitcs.2012.08.03.

[2] Anastassiou, D.: Genomic signal processing. *IEEE Signal Processing Magazine*, ročník 18, č. 4, 2001: s. 8–20, ISSN 10535888, doi:10.1109/79.939833.

[3] Arniker, S. B.; Kwan, H. K.: Graphical representation of DNA sequences. *2009 IEEE International Conference on Electro/Information Technology*, Červen 2009: s. 311–314, doi:10.1109/EIT.2009.5189633.

[4] Baxevanis, A.; Ouellette, B.: *Bioinformatics: a practical guide to the analysis of genes and proteins*. New York, New York, USA: John Wiley & Sons, Inc., druhé vydání, 2004, ISBN 0471383902, 495 s.

[5] Berger, J. a.; Mitra, S. K.; Carli, M.; aj.: Visualization and analysis of DNA sequences using DNA walks. *Journal of the Franklin Institute*, ročník 341, č. 1-2, Leden 2004: s. 37–53, ISSN 00160032, doi:10.1016/j.jfranklin.2003.12.002.

[6] Boore, J. L.: Animal mitochondrial genomes. *Nucleic acids research*, ročník 27, č. 8, Duben 1999: s. 1767–80, ISSN 0305-1048.

[7] Brinkman, F. S.; Leipe, D. D.: Phylogenetic analysis. *Methods of biochemical analysis*, ročník 43, Leden 2001: s. 323–58, ISSN 0076-6941.

[8] Chapdelaine, Y.; Bonen, L.: The wheat mitochondrial gene for subunit I of the NADH dehydrogenase complex: a trans-splicing model for this gene-in-pieces. *Cell*, ročník 65, č. 3, Květen 1991: s. 465–72, ISSN 0092-8674.

[9] Crick, F.: Central dogma of molecular biology. *Nature*, 1970.

[10] Cristea, P.: Phase analysis of DNA genomic signals. *Proceedings of the 2003 International Symposium on Circuits and Systems, 2003. ISCAS '03.*, ročník 5, 2003: s. V–25–V–28, doi:10.1109/ISCAS.2003.1206163.

[11] Cristea, P. D.: Conversion of nucleotides sequences into genomic signals. *Journal of cellular and molecular medicine*, ročník 6, č. 2, 2002: s. 279–303, ISSN 1582-1838.

[12] Cristea, P. D.: Large scale features in DNA genomic signals. *Signal Processing*, ročník 83, č. 4, Duben 2003: s. 871–888, ISSN 01651684, doi:10.1016/S0165-1684(02)00477-2.

[13] Cristea, P. D.; Tuduce, R.: Signal processing of genomic information: mitochondrial genomic signals of hominidae. *Proceedings ECVIPMC 2003 4th EURASIP Conference focused on VideoImage Processing and Multimedia Communications IEEE Cat No03EX667*, ročník 1, 2003, doi:10.1109/VIPMC.2003.1220463.

[14] Cvrčková, F.: *Úvod do praktické bioinformatiky*. Praha: Academia, vyd. 1. vydání, 2006, ISBN 80-200-1360-1, 148 s.

[15] Gray, M.: Mitochondrial evolution. *Cold Spring Harbor Perspectives in Biology*, ročník 4, č. 9, Září 2012: str. a011403, ISSN 1943-0264, doi:10.1101/cshperspect.a011403.

[16] Hamori, E.; Ruskin, J.: H curves, a novel method of representation of nucleotide series especially suited for long DNA sequences. *Journal of Biological Chemistry*, 1983.

[17] Hanuš, J.: *Shluková analýza a její aplikace*. Bakalárská práce, Západočeská univerzita v Plzni, 2009.
URL `<https://stag-ws.zcu.cz/ws/services/rest/kvalifikacniprace/downloadPraceContent&adipIdno=31640>`

[18] Keeling, P. J.; Archibald, J. M.: Organelle evolution: what's in a name? *Current biology : CB*, ročník 18, č. 8, Duben 2008: s. R345–7, ISSN 0960-9822, doi:10.1016/j.cub.2008.02.065.

[19] Khalimonchuk, O.; Rödel, G.: Biogenesis of cytochrome c oxidase. *Mitochondrion*, ročník 5, č. 6, Prosinec 2005: s. 363–88, ISSN 1567-7249, doi:10.1016/j.mito.2005.08.002.

[20] Knoop, V.: The mitochondrial DNA of land plants: peculiarities in phylogenetic perspective. *Current genetics*, ročník 46, č. 3, Září 2004: s. 123–39, ISSN 0172-8083, doi:10.1007/s00294-004-0522-8.

[21] Kozumplík, J.: Umělá inteligence v medicíně : Shluková analýza, 2013.
URL `<https://www.vutbr.cz/elearning/file.php/111771/prednasky/AUIN\_04\_shlukova\_analyza.>`

[22] Kuang, D.; Wu, H.; Wang, Y.; aj.: Complete chloroplast genome sequence of Magnolia kwangsiensis (Magnoliaceae): implication for DNA barcoding and population genetics. . . . , ročník 673, 2011: s. 663–673, doi:10.1139/G11-026.

[23] Kwan, H. K.; Arniker, S. B.: Numerical representation of DNA sequences. *2009 IEEE International Conference on Electro/Information Technology*, Červen 2009: s. 307–310, doi:10.1109/EIT.2009.5189632.

[24] Lemire, B.: Mitochondrial genetics. *WormBook : the online review of C. elegans biology*, Leden 2005: s. 1–10, ISSN 1551-8507, doi:10.1895/wormbook.1.25.1.

[25] Ma, X.-Y.; Xie, C.-X.; Liu, C.; aj.: Species identification of medicinal pteridophytes by a DNA barcode marker, the chloroplast psbA-trnH intergenic region. *Biological & pharmaceutical bulletin*, ročník 33, č. 11, Leden 2010: s. 1919–24, ISSN 1347-5215.

[26] Mach, J.: Cool as the Cucumber Mitochondrial Genome: Complete Sequencing Reveals Dynamics of Recombination, Sequence Transfer, and Multichromosomal Structure. *The Plant Cell*, ročník 23, č. 7, 2011: s. 2472–2472, ISSN 1040-4651, doi:10.1105/tpc.111.230711.

[27] Marienfeld, J. R.; Newton, K. J.: The nad4 gene of maize mitochondria is highly conserved. *Plant physiology*, ročník 104, č. 1, Leden 1994: s. 301–2, ISSN 0032-0889.

[28] Martin, W. F.; Mentel, M.: The Origin of Mitochondria. 2010.
URL <http://www.nature.com/scitable/topicpage/the-origin-of-mitochondria-14232356>

[29] Nair, C.: Mitochondrial genome organization and cytoplasmic male sterility in plants. *Journal of biosciences*, ročník 18, č. 3, 1993: s. 407–422.

[30] Nečas, O.; Svoboda, A.; Hejtmánek, M.; aj.: *Obecná biologie pro lékarské fakulty*. Nakladatelství H+H, Jinočany, třetí vydání, 2000, ISBN 80-86022-46-3, 555 s.

[31] Organization, B.: Scoring matrix - Bioinformatics.Org Wiki. 2007.
URL <http://www.bioinformatics.org/wiki/Scoring_matrix>

[32] Palmer, J. D.; Herbon, L. a.: Plant mitochondrial DNA evolves rapidly in structure, but slowly in sequence. *Journal of molecular evolution*, ročník 28, č. 1-2, 1988: s. 87–97, ISSN 0022-2844.

[33] Peng, C.; Buldyrev, S.; Havlin, S.: Mosaic organization of DNA nucleotides. *Physical Review E*, 1994.

[34] Randić, M.; Vracko, M.; Nandy, A.; aj.: On 3-D graphical representation of DNA primary sequences and their numerical characterization. *Journal of chemical*

information and computer sciences, ročník 40, č. 5, 2000: s. 1235–44, ISSN 0095-2338.

[35] Rotterová, J.: *Metoda DNA barcodingu a její využití u protist*. Bakalárská práce, Univerzita Karlova v Praze, 2012.
URL `<https://is.cuni.cz/webapps/zzp/detail/116819/>`

[36] Roy, A.; Raychaudhury, C.; Nandy, A.: Novel techniques of graphical representation and analysis of DNA sequencesâ€"a review. *Journal of Biosciences*, , č. 1, 1998: s. 55–71.

[37] Sakoe, H.; Chiba, S.: Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ročník 26, č. 1, Únor 1978: s. 43–49, ISSN 0096-3518, doi: 10.1109/TASSP.1978.1163055.

[38] Skutkova, H.; Vitek, M.; Babula, P.; aj.: Classification of genomic signals using dynamic time warping. *BMC Bioinformatics*, ročník 14, č. Suppl 10, 2013: str. S1, ISSN 1471-2105, doi:10.1186/1471-2105-14-S10-S1.

[39] Sloan, D. B.: One ring to rule them all? Genome sequencing provides new insights into the â€master circleâ€™ model of plant mitochondrial DNA structure. *New Phytologist*, ročník 200, č. 4, Prosinec 2013: s. 978–985, ISSN 0028646X, doi:10.1111/nph.12395.

[40] Stern, D. B.; Palmer, J. D.: Extensive and widespread homologies between mitochondrial DNA and chloroplast DNA in plants. *Proceedings of the National Academy of Sciences of the United States of America*, ročník 81, č. 7, Duben 1984: s. 1946–50, ISSN 0027-8424.

[41] Voss, R.: Evolution of long-range fractal correlations and 1/f noise in DNA base sequences. *Physical review letters*, ročník 68, č. 25, 1992: s. 3805–3808.

[42] Watson, J.; Crick, F.: A structure for deoxyribose nucleic acid. *Nature*, 1953.

[43] Wöllmer, M.; Al-Hames, M.; Eyben, F.; aj.: A multidimensional dynamic time warping algorithm for efficient multimodal fusion of asynchronous data streams. *Neurocomputing*, 2009.

[44] Yoon, H. S.; Reyes-Prieto, A.; Melkonian, M.; aj.: Minimal plastid genome evolution in the Paulinella endosymbiont. *Current biology : CB*, ročník 16, č. 17, Září 2006: s. R670–2, ISSN 0960-9822, doi:10.1016/j.cub.2006.08.018.

[45] Zhang, R.; Zhang, C.-T.: Identification of replication origins in archaeal genomes based on the Z-curve method. *Archaea (Vancouver, B.C.)*, ročník 1, č. 5, Květen 2005: s. 335–46, ISSN 1472-3646.

[46] Zhang, Z.-J.: DV-Curve: a novel intuitive tool for visualizing and analyzing DNA sequences. *Bioinformatics (Oxford, England)*, ročník 25, č. 9, Květen 2009: s. 1112–7, ISSN 1367-4811, doi:10.1093/bioinformatics/btp130.

SEZNAM SYMBOLŮ, VELIČIN A ZKRATEK

IUPAC	International Union of Pure and Applied Chemistry
DNA	kyselina deoxyribonukleová
cpDNA	chloroplastová DNA
mtDNA	mitochondriální DNA
bp	base pair, pár bazí
NCBI	National Center for Biotechnology Information
DTW	dynamické borcení časové osy
DSP	číslicové zpracování signálu
UPGMA	Unweighted Pair Group Method with Arithmetic Mean
RNA	Ribonukleová kyselina
mRNA	mediátorová RNA
tRNA	transferová RNA
rRNA	ribozomální RNA

SEZNAM PŘÍLOH

A PŘÍLOHA A

A.1 Přehled sekvencí mtDNA z NCBI

Tab. A.1: Údaje z databáze NCBI ke genetickému materiálu mtDNA

Latinský název	**ID NCBI**	**Délka genomu bp**
Amborella trichopoda	KF754803.1,KF754801.1	3179272 bp,187116 bp
Arabidopsis thaliana	NC_001284.2	366924 bp
Daucus carota	JQ248574.1	281132 bp
Glycine max	NC_020455.1	402558 bp
Chaetosphaeridium globosum	NC_004118.1	56574 bp
Chara vulgaris	NC_005255.1	67737 bp
Chlamydomonas reinhardtii	NC_001638.1	15758 bp
Chlorokybus atmophyticus	NC_009630.1	201763 bp
Liriodendron tulipifera	NC_021152.1	553721 bp
Lotus japonicus	NC_016743.2	380861 bp
Mesostigma viride	NC_008240.1	42424 bp
Nicotiana tabacum	NC_006581.1	430597 bp
Oryza sativa	NC_011033.1	490520 bp
Ostreococcus tauri	NC_008290.1	44237 bp
Physcomitrella patens	NC_007945.1	105340 bp
Pseudendoclonium akinetum	AY359242.1	95880 bp
Pyropia yezoensis	NC_017837.1	41688 bp
Sorghum bicolor	NC_008360.1	468628 bp
Vitis vinifera	NC_012119.1	773279 bp
Zea mays	AY506529.1	569630 bp

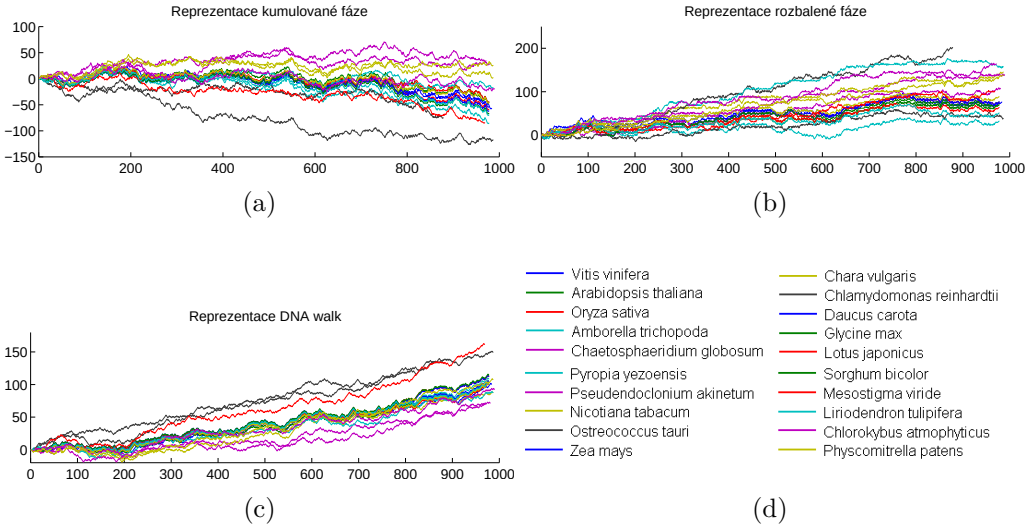
A.2 Přehled sekvencí cpDNA z NCBI

Tab. A.2: Údaje z databáze NCBI ke genetickému materiálu cpDNA

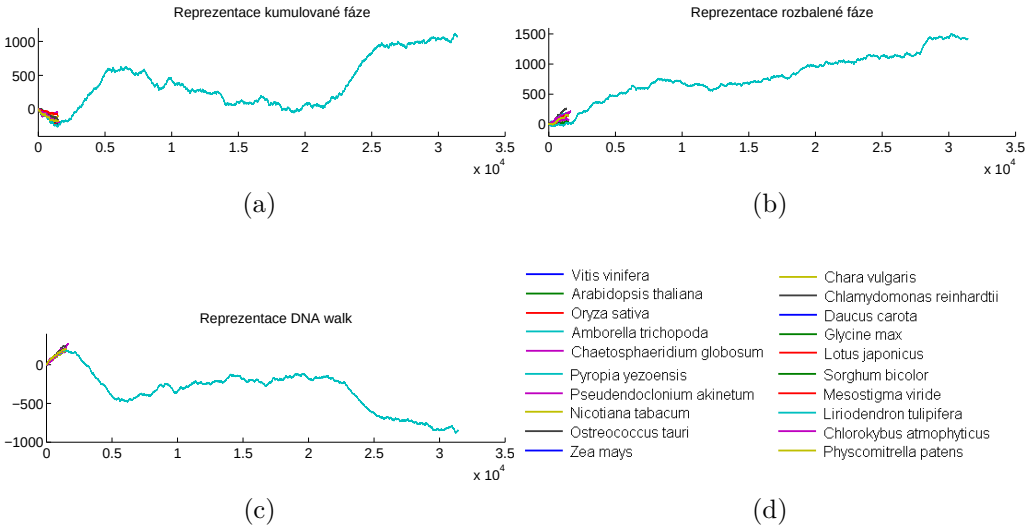
Latinský název	**ID NCBI**	**Délka genomu bp**
Amborella trichopoda	AJ506156.2	162686 bp
Arabidopsis thaliana	AP000423.1	154478 bp
Daucus carota	DQ898156.1	155911 bp
Glycine max	DQ317523.1	152218 bp
Chaetosphaeridium globosum	AF494278.1	131183 bp
Chara vulgaris	DQ229107.1	184933 bp
Chlamydomonas reinhardtii	BK000554.2	203828 bp
Chlorokybus atmophyticus	DQ422812.2	152254 bp
Liriodendron tulipifera	DQ899947.1	159886 bp
Lotus japonicus	AP002983.1	150519 bp
Mesostigma viride	AF166114.1	118360 bp
Nicotiana tabacum	Z00044.2	155943 bp
Oryza sativa	NC_001320.1	134525 bp
Ostreococcus tauri	CR954199.2	71666 bp
Physcomitrella patens	AP005672.1	122890 bp
Pseudendoclonium akinetum	AY835431.1	195867 bp
Pyropia yezoensis	KC517072.1	191975 bp
Sorghum bicolor	EF115542.1	140754 bp
Vitis vinifera	NC_007957.1	160928 bp
Zea mays	NC_001666.2	140384 bp

B PŘÍLOHA B

B.1 Numerické reprezentace genů nad1, nad4

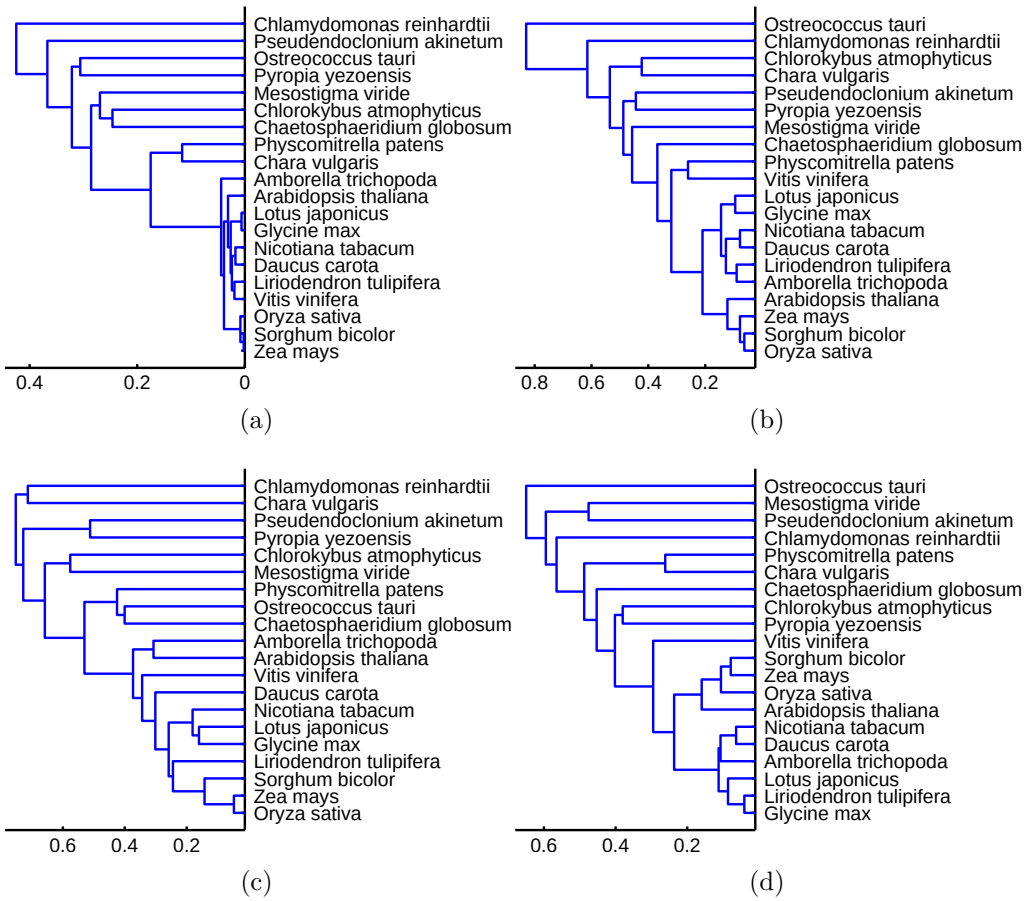


Obr. B.1: Průběhy numerických sekvencí DNA genu nad1

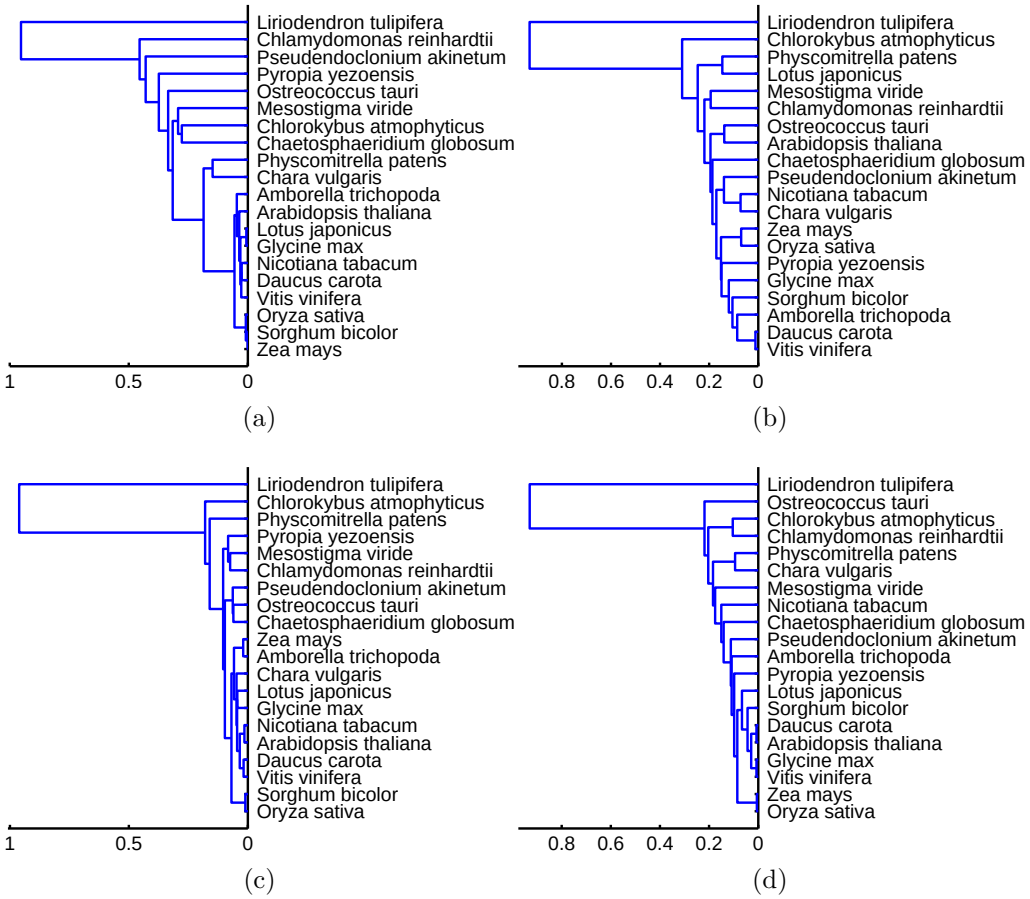


Obr. B.2: Průběhy numerických sekvencí DNA genu nad4

B.2 Dendrogramy genů nad1, nad4

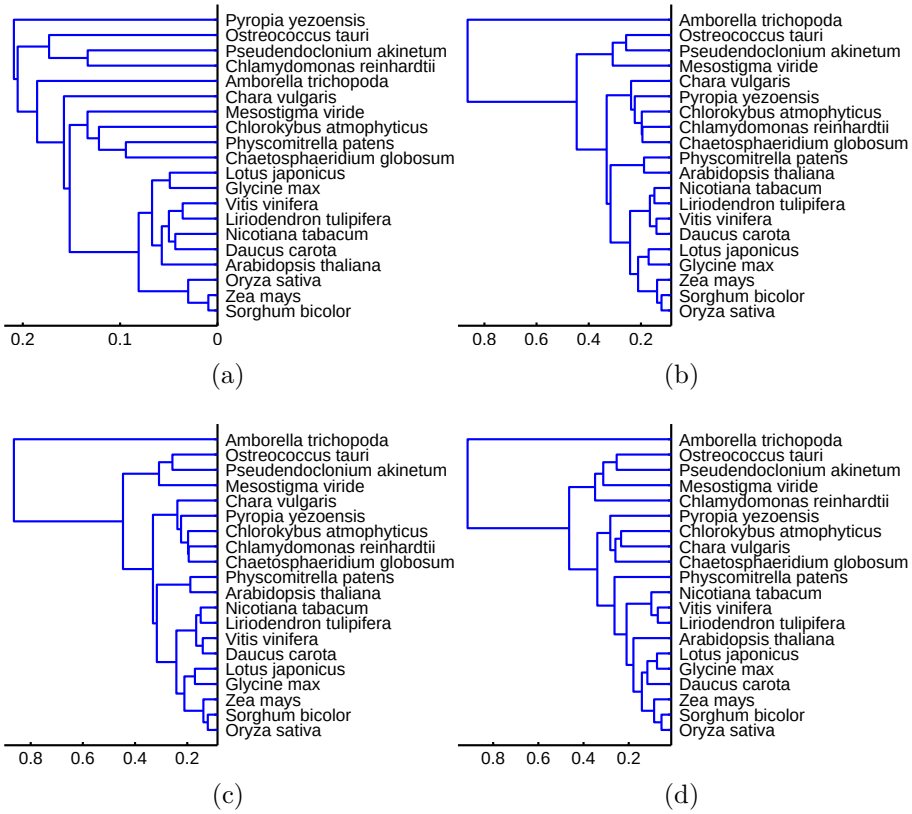


Obr. B.3: Dendrogramy genu nad1 a)Znakové metody (globální zar.) b)Kumulované f. c)Rozbalené f. d)DNA walk

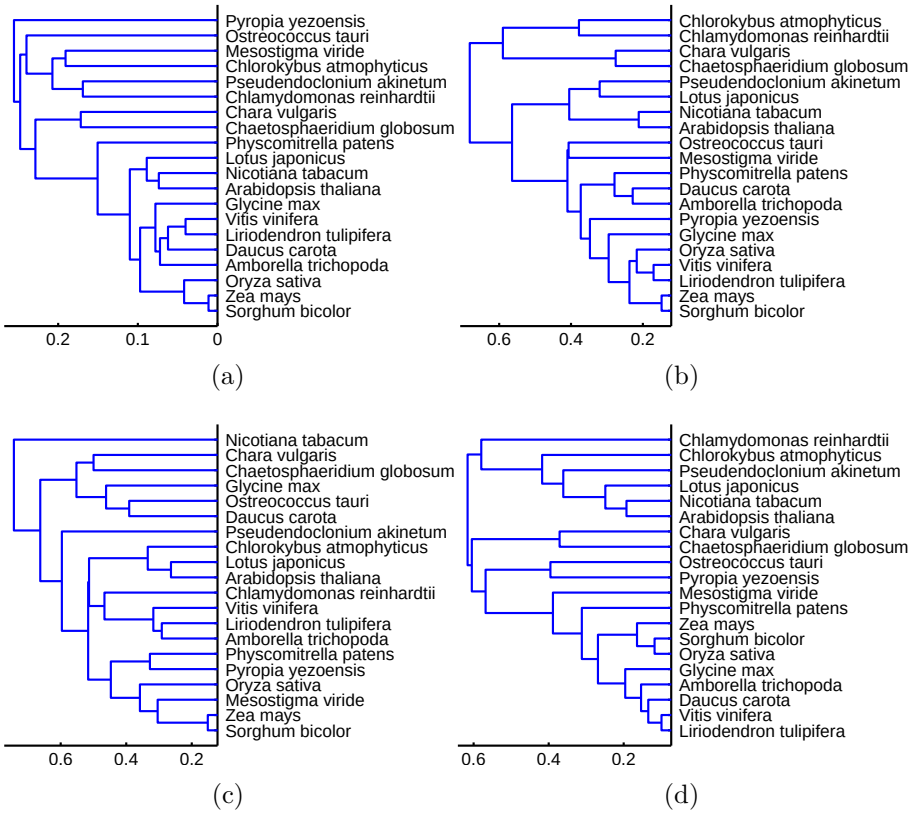


Obr. B.4: Dendrogramy genu nad4 a)Znakové metody (globální zar.) b)Kumulované f. c)Rozbalené f. d)DNA walk

B.3 Dendrogramy genů psbA, psbC



Obr. B.5: Dendrogramy ze sekvencí - psbA a)Znakové metody b)Kumulované f. c)Rozbalené f. d)DNA walk



Obr. B.6: Dendrogramy ze sekvencí - psbC a)Znakové metody b)Kumulované f. c)Rozbalené f. d)DNA walk

C PŘÍLOHA C

C.1 Obsah přiloženého CD

- Text práce ve formátu pdf
- KlasifikaceGUI - spustitelná aplikace
- KlasifikaceGUI - zdrojové kódy
- Ukázkové sekvence ve formátu fasta
- Nápověda k aplikaci KlasifikaceGUI